UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL


# *People Analytics - Flight Risk*

Ana Catarina Costa Dias Queiroz

**Mestrado em Matemática Aplicada à Economia e Gestão**

Trabalho de Projeto orientado por:
Prof.ª Doutora Raquel João Fonseca
Prof. Doutor João Telhada

2020

*"You won't lose your job to a computer, you'll lose it to a human who is better at using a computer"*

Stephane Scheyven, *Contagious Communications*

# Resumo

Nos últimos anos, tem-se vindo a registar um aumento da rotatividade laboral. Efectuar uma avaliação adequada aos riscos associados à rotatividade dos colaboradores, pode contribuir para a diminuição dos custos associados às novas contratações e evitar a perda de produtividade da organização. Nesse sentido, com este trabalho pretende-se, como objectivo global, criar um modelo capaz de prever antecipadamente as saídas voluntárias dos colaboradores, com o recurso à metodologia de *Human Resources Analytics* (HRA), que corresponde ao conjunto de competências, tecnologias e práticas que permite aos Recursos Humanos (RH), com base na exploração de dados, fornecer *insights* de suporte à tomada de decisão, na gestão e resolução de desafios de negócio. Como tal, pretende-se criar dois modelos distintos de previsão de saída dos colaboradores, através do uso de modelos de regressão logística e de árvores de decisão. Estas metodologias permitem, através de um conjunto de variáveis independentes, prever a rescisão voluntária do contrato de trabalho. Para além disso, é possível averiguar qual é a variável que tem o maior ganho de informação no modelo. Para a construção dos modelos, a variável resposta é definida como a vinculação, ou não, de um colaborador à empresa, consoante as variáveis que melhor caracterizam o seu perfil.

Por neste trabalho se ter assumido o propósito de identificar e quantificar os colaboradores que pretendem rescindir o contrato de trabalho, está-se perante a um modelo de classificação. Desta forma, são estabelecidas duas amostras distintas, de treino e de teste. Assim, o conjunto de observações da amostra de teste nunca irá influenciar a construção do modelo e, por sua vez, é possível testar e avaliar a capacidade discriminatória do mesmo. O melhor modelo obtido foi através do uso da regressão logística, que permitiu predizer correctamente todas as observações em 74,71%, sendo a variável idade aquela que tem maior importância no modelo.

**Palavras-chave:** Risco de Saída, *Human Resources Analytics*, Modelo de Regressão Logística; Árvores de Decisão; *Employee Turnover*.

# Abstract

In recent years, there has been an increase in labour turnover. Carrying out an adequate assessment of the risks associated with employee turnover may contribute to the reduction of costs associated with new hires and avoid the loss of productivity of the organisation. In this sense, the overall objective of this work is to create a model capable of predicting in advance the voluntary departures of employees, using the Human Resources Analytics methodology, which corresponds to the set of skills, technologies and practices that allow Human Resources (HR), based on data exploration, to provide insights to support decision making, management and resolution of business challenges. As such, it aims to create two distinct models of employee exit forecasting, through the use of logistic regression models and decision trees. These methodologies allow, through a set of independent variables, the voluntary termination of the employment contract. In addition, it is possible to find out which variable has the greatest information gain in the model. For the construction of the models, the response variable is defined as the link, or not, of an employee to the company, depending on the variables that best characterize his/her profile.

Since the purpose of this work is to identify and quantify the employees who intend to terminate the employment contract, this is a classification model. In this way, two distinct training and test samples are established. Thus, the set of observations in the test sample will never influence the construction of the model and, in turn, it is possible to test and evaluate its discriminatory capacity. The best model obtained was through the use of logistic regression, which allowed the correct prediction of all observations at 74.71%, the age variable being the one that has the greatest importance in the model.

**Keywords:** Flight Risk; Human Resources Analytics; Logistic Regression Model; Decision Tree; Employee Turnover.

# Agradecimentos

Dedico este trabalho à minha mãe e ao meu irmão pelo apoio incondicional que me deram durante todo o meu percurso, não só no carácter académico, como ao longo da minha vida. Aproveito para dizer que são vocês que me dão protecção, força, que me ensinam a importância da união e, como tal, não posso estar mais grata por vos ter comigo.

À minha mãe, Mafalda Queiroz, agradeço tudo o que me ensinou sobre a vida, que apesar de todas as adversidades que tivemos que enfrentar, nunca me deixou de desistir dos meus sonhos. Acima de tudo, ensinou-me que é possível ser mãe e pai ao mesmo tempo e, como tal, mostrou-me a importância de permanecermos juntas.

Ao meu irmão, Filipe Queiroz, que me incutiu o gosto pela matemática, agradeço todo o amor que demonstra por mim nos mais bonitos e simples gestos. A ele, agradeço todas as horas que se sentou ao meu lado a ensinar cada detalhe da vida para que eu crescesse tanto a nível profissional e, acima de tudo, a nível pessoal. A ti, すべてに感謝します.

Aos meus avós maternos (*in memoriam*), Maria Isabel Queiroz e José António Queiroz, uma menção especial e um amor incondicional eterno. Saudades vossas.

Um agradecimento especial aos meus orientadores, professora Raquel João Fonseca e professor João Telhada, por serem sempre tão disponíveis, dedicados e pacientes. Obrigada por todas as críticas e conselhos que simplificaram as adversidades que surgiram.

Aos meus professores de mestrado, em especial à professora Teresa Alpuim, pelo seu tempo, ajuda e interesse ao longo do desenvolvimento deste projecto. Aproveito para agradecer a todos os professores do Departamento de Estatística e Investigação Operacional, da Faculdade de Ciências da Universidade de Lisboa, que contribuíram para a minha formação académica transmitindo conhecimento e valores, permitindo a valorização deste trabalho.

À Alexandra Almeida, agradeço toda a amizade que temos, por todo o carinho, por tantas horas de gargalhada e por tudo aquilo que ainda vamos viver juntas. Permitiste que o meu percurso académico não se resumisse a uma caminhada sozinha e, devo-te todos os minutos de alegria, de choro e de companhia. Na verdade, há momentos que mereciam um *replay*. Este trabalho também é teu.

À Joana Canilho, por ter sido uma excelente surpresa durante o percurso do mestrado. Agradeço todos os ensinamentos que me transmitiu e, principalmente mostrou-me que temos que seguir os nossos próprios sonhos, para que possamos ser ainda mais felizes.

Um enorme obrigada aos meus colegas de mestrado, aos meus melhores amigos, que apesar de toda a distância, conseguiram acompanhar e contribuir para o meu percurso académico e pessoal.

Agradeço aos meus colegas Anabela Paulino, Filipa Marques, Diogo Tavares Antunes, Mónica Fernandes e Luís Moura por terem estado tão presentes, por toda a paciência e celebrações de cada conquista. Um agradecimento especial ao Carlos Graça, por me desafiar a arriscar e por me mostrar sempre o lado bonito das coisas.

Um agradecimento inevitável aos meus colegas e amigos da WTW, Vânia Vieira, Rúben Lavrador,

Ana Sousa e Lara Anes, por toda a compreensão, paciência e dedicação.

Um enorme obrigada a todos, por confiarem, não desistirem de mim e por vibrarem comigo cada conquista.

Fevereiro de 2020,
Catarina Queiroz

# Índice

**ÍNDICE**

# Lista de Figuras

# Lista de Tabelas

# Lista de Acrónimos e Siglas

**AIC** Critério de Informação de *Akaike*/*Akaike Information Criterion*
**AUC** Área Abaixo da Curva/*Area Under the Curve*

**DDD** *Data Driven Decision*

**FN** Falsos Negativos
**FP** Falsos Positivos

**GRH** Gestão de Recursos Humanos

**HRA** *Human Resources Analytics*

**KDD** *Knowledge Discovery in Databases*

**OR** *Odds Ratio*

**PA** *People Analytics*
**PSI20** *Portuguese Stock Index*

**RGPD** Regulamento Geral sobre a Protecção de Dados
**RH** Recursos Humanos
**ROC** *Receiver Operating Characteristic*

**VIF** *Variance Inflation Factor*
**VN** Verdadeiros Negativos
**VP** Verdadeiros Positivos

# 1. Enquadramento

Com o crescimento tecnológico exponencial que se registou nas últimas décadas, é cada vez mais fácil capturar e armazenar dados de negócio. Na verdade, é praticamente impossível encontrar actualmente uma empresa de média ou de grande dimensão que não use sistemas avançados de informação, de forma a melhorar a gestão do seu negócio. No entanto, este aumento significativo na facilidade em obter e armazenar grandes volumes de dados, não foi devidamente acompanhado pela capacidade para interpretar esses mesmos dados (Smyth et al., 1996). Apesar destes serem armazenados e, de certa forma, bem estruturados e orientados ao negócio, não são, na grande maioria dos casos, de fácil interpretação quando são utilizados meios de análise convencionais. É portanto comum existirem casos que constituem um grande volume de dados, com informação válida, que poderá futuramente possuir um grande valor comercial, mas que no entanto não se sabe como extrair e gerar desses dados alguma informação válida e útil para as actividades empresariais. As situações referidas anteriormente possuem frequentemente no seu histórico de dados, padrões implícitos que se encontram por descobrir e que, por sua vez, poderão conter informações bastante relevantes para o negócio. É neste contexto que surge em 1989 o conceito de *Knowledge Discovery in Databases* (KDD).

O KDD surge como um ramo da computação, que tem como principal objectivo a extracção de conhecimento útil a partir dos dados, conhecimento este que não seria capaz de ser detectado usando as técnicas de análise adoptadas até ao momento em que surgiu esse algoritmo. Conceito idêntico, o *data mining*, não deve ser confundido com o processo de KDD, uma vez que, representa uma das fases do KDD. O *data mining* é definido como sendo a aplicação de determinados algoritmos para a extracções de padrões, enquanto que o *KDD* é visto como o processo de descoberta de informação útil a partir dos dados (Smyth et al., 1996 e Gomes, 2011).

## 1.1 *People Analytics*

Nos últimos anos, a Gestão de Recursos Humanos (GRH) tem vindo a modificar-se, passando de uma área operacional para uma área estratégica e, consequentemente, tornou-se num sector empresarial mais atractivo. O *Human Resources Analytics* (HRA) é um tema que está a crescer de forma exponencial entre os profissionais de Recursos Humanos. A análise preditiva é uma tendência futura na área de RH. As ferramentas de recrutamento prevêem os melhores desempenhos e, cada vez mais as empresas são capazes de prever qual é o colaborador que provavelmente irá sair da empresa de forma voluntária. Num mercado de trabalho cada vez mais complexo e concorrencial, a corrida pela vantagem competitiva centra-se, hoje, na compreensão de todos os elementos que caracterizam a força de trabalho. O HRA é o conjunto de competências, tecnologias e práticas que permite aos RH, com base na exploração de dados, fornecer *insights* de suporte à tomada de decisão, bem como a gestão e resolução de desafios de negócios. Isto é, o HRA tem por base a utilização de uma abordagem multidimensional orientada a dados para suportar as decisões em torno das práticas, programas e processos de gestão de pessoas.

## 1. ENQUADRAMENTO

Segundo Brynjolfsson et al., 2014, o desempenho geral de uma entidade empresarial é superior em empresas que evidenciam o uso de decisões na gestão do negócio, com base em análises de dados, designado por *Data Driven Decision* (DDD). A 179 grandes empresas foi efectuada uma recolha de dados sobre as práticas comerciais e os investimentos efectuados em tecnologia. Desta amostra, as empresas que adoptam o método DDD, a fim de tomar as melhores decisões a nível do negócio, são 5 a 6% mais produtivas. O uso do DDD também está fortemente associado a um nível significativamente mais elevado de rentabilidade e de valor de mercado (Bersin, 2014).

Actualmente, é visível a importância da utilização dos dados nas empresas. No entanto, é necessário adquirir competências para saber recolher informações relevantes e accionáveis. Esta recolha é necessário que seja o mais eficaz possível, de forma a que a competitividade empresarial seja cada vez mais interessante ao nível do negócio. Para tal, é importante perceber quais são as melhores metodologias a utilizar e, por sua vez, quais são as diferenças nos resultados da empresa.

Em 2013, o HRO Today Institute investigou pela primeira vez este tema e concluiu que as organizações que usam o HRA para optimizar o recrutamento de talentos superam a concorrência empresarial em 58% do tempo do processo de recrutamento (Institute, 2015).

Posto isto, existem vários caminhos para criar uma função de análise bem sucedida. A obtenção de recursos necessários para entender e interpretar os dados e sistemas da organização e, por sua vez, identificar problemas de qualidade dos dados torna-se num processo bastante exaustivo e moroso. Este processo traduz-se numa área de investimento contínuo à medida que o mercado analista evolui.

A Bersin by Deloitte efectuou um estudo e concluiu que 86% das organizações da amostra recolhida, estão fortemente focadas em relatórios extensos e não potencializam o poder das análises que podem efectuar, baseando-se apenas em análises convencionais. A grande maioria destas organizações produzem métricas para fins de conformidade e operam de modo reactivo, isto é, produzir resultados atendendo apenas a solicitações de mercado. Uma pequena parte da amostra adopta uma abordagem pro-activa, utilizando tendências de negócio de forma a destacar o que está a funcionar, ou não, no âmbito da sua organização. Contudo, apenas 10% das organizações deste estudo se situam numa posição mais avançada, devido à utilização de análises mais complexas, de forma a ajudar os líderes de negócio a resolver desafios de recrutamento de talento. Por fim, apenas 4% da amostra utiliza a análise preditiva para perceber o comportamento futuro dos talentos que visam recrutar (Deloitte, 2013).

A Deloitte desenvolveu um modelo de maturidade consoante a utilização de métricas analíticas. Assim, cada organização é capaz de deduzir os recursos necessários para aumentar a produtividade e a utilização de análises estatísticas mais complexas (tabela 1.1).

Tabela 1.1: *Talent Analytics Maturity Model (Deloitte).*

| **Nível** | **Competências** |
|---|---|
| 1. *Operational Reporting* | 1.a. Relatórios operacionais e reactivos focados em medidas de eficiência e conformidade;<br>1.b. foco na precisão, consistência e coerência dos dados. |
| 2. *Advanced Reporting* | 2.a. Relatórios operacionais e pro-activos para *benchmarking* e tomada de decisão;<br>2.b. análise multidimensional. |
| 3. *Advanced Analytics* | 3.a. Modelagem estatística e análise da principal causa para |

| **Nível** | **Competências** |
|---|---|
| | resolver problemas de negócios; |
| | 3.b. identificar pro-activamente problemas; |
| | 3.c. recomendar soluções accionáveis. |
| 4. *Predictive Analytics* | 4.a. Desenvolvimento de modelos preditivos; |
| | 4.b. criação de cenários; |
| | 4.c. análise e mitigação de riscos; |
| | 4.d. integração com planeamento estratégico. |

As diferenças existentes entre as organizações desta amostra demonstram que é possível distinguirem-se em diferentes níveis de maturidade, face à utilização de HRA. Por sua vez, esta divisão permitiu definir diversos factores associados à maturidade da utilização de métricas analíticas avançadas de uma organização. Entre os diferentes factores definidos, destacam-se (Deloitte, 2013):

- *Strong Technical Skills* - as organizações dos níveis 3 e 4 do modelo de maturidade desenvolvem fortes habilidades estatísticas e de compreensão de dados. Mais de 70% das organizações do nível 4 são compostas por colaboradores com experiência em análise de dados, bases de dados e visualização dos mesmos;

- *Beyond Number Crunching* - corresponde à capacidade de compreender e criar funções analíticas complexas, com o intuito de resolver problemas do ponto de vista do negócio. Uma das maiores dificuldades das equipas de *analytics* resume-se à falta de capacidade de "contar uma história" por detrás dos dados, para que os gestores de negócio possam entender rapidamente as implicações e as acções a tomar. Para tal, é necessário criar um meio ambiente dinâmico entre diversas áreas, para que seja possível potencializar as habilidades de cada colaborador, nomeadamente das áreas de RH e consultoria;

- *Data Quality* - as estruturas organizativas que pretendem atingir uma maturidade a nível analítico precisam de garantir que os dados são de elevada qualidade e, preferencialmente, actualizados.

- *Effective Dashboards* - as equipas de elevada maturidade analítica criam plataformas digitais, para que seja possível aceder mais rapidamente às informações necessárias, a fim de gerar *insights* e recomendações sobre os desafios do negócio.

O caminho para a obtenção dos níveis elevados de maturidade analítica pode não ser fácil. Contudo, o esforço necessário da organização resulta em ganhos significativos de eficiência. As organizações que atingem a maturidade analítica são capazes de, detalhadamente:

- em média, obter o dobro da probabilidade de melhorar o recrutamento;

- alcançar o triplo da eficiência;

- aumentar em 2,5 vezes a retenção de talento, através de mobilidade interna.

Desta forma, é necessário saber como se pode tornar o HRA num ecossistema analítico. Assim, surge o conceito de *People Analytics* (PA), que consiste num conjunto de processos, facilitados por tecnologia, que tira partido de métodos descritivos, visuais e estatísticos para interpretar dados de pessoas e processos de RH (Marler e Boudreau, 2016). Traduz-se numa abordagem baseada em evidências para tomar melhores decisões na gestão de capital humano, assente num conjunto de ferramentas e tecnologias, que varia do *report* simples de métricas elementares à modelação preditiva (Madsen e Slåtten, 2017).


Figura 1.1: Processo para a obtenção do ecossistema analítico.

Através da figura 1.1 é possível verificar que existem sete etapas para a obtenção do ecossistema analítico (Empresa, 2017).

A etapa definida como visão caracteriza-se por garantir que todas as decisões de gestão do capital humano estão assentes em dados e análises. Através de uma cultura de dados e análises de apoio à tomada de decisão, pretende-se optimizar os resultados da gestão de pessoas e promover o *empowerment* dos colaboradores.

Sabe-se que a etapa da organização foi bem sucedida se for possível: adaptar a proposta de valor às expectativas e necessidades individuais; implementar uma abordagem consistente de inovação nas práticas de RH, que permita compreender os sucessos e insucessos das acções de *engagement* e resultados de desenvolvimento e retenção de talento e, por fim, estabelecer uma óptima experiência do colaborador.

A etapa da organização caracteriza-se pela capacidade do uso de HRA tirar partido de um conjunto de técnicas analíticas desde *reporting* descritivo a análises preditivas, mas também estudos experimentais que possam apoiar a concretização de novos *insights*, solucionar problemas e orientar as acções de RH.

Tabela 1.2: Modelo da etapa organização.

| **Estágio** | | **Negócio** | ***HR Analytics*** | **IT** |
|---|---|---|---|---|
| I. Problema | | Identifica | Identifica/Prioriza | - |
| II. Solução | a. Recolha de dados | - | Identifica e Processa | Suporta |
| | b. Análise de dados | - | Modela e Reporta | - |
| | c. Interpretação | Apoia | Gera *insights* | - |
| III. Acção | | Facilita | Gera a mudança | Apoia |
| Acesso à ferramenta HRA | | Solicita | Decide | Apoia |

Através da tabela 1.2 é possível verificar que esta etapa está dividida por 3 estágios, nomeadamente, o problema, a solução e a acção. Pode-se verificar que, a identificação do problema cabe à área de negócio que, consequentemente, comunica ao HRA. No entanto, qualquer tipo de problema identificado tem que

ser priorizado em função dos restantes problemas já identificados (Empresa, 2017).

O segundo estágio é dividido em três importantes etapas, a recolha, a análise e a interpretação de dados. A recolha dos dados compete às funções de HRA identificar e processar, suportada pelas áreas de IT, isto é, tecnologia e informação. Da mesma forma, a análise de dados é modelada e reportada através do HRA. Já a interpretação da análise de dados é apoiada através de especialistas em áreas de negócio, de maneira a que seja possível gerar recomendações accionáveis.

Relativamente ao último estágio, as acções pressupõem a mudança na organização, acabando por ser apoiadas pela área de IT e, consequentemente, a área de negócio patrocina as acções.

O acesso à ferramenta HRA é solicitado pela área de negócio, tendo por base a decisão do HRA e apoiada pelas áreas de IT, através da distribuição e configuração de licenças para o acesso à mesma.

O sucesso do HRA dependerá, assim, da capacidade de tirar partido de um conjunto equilibrado de competências para a entrega efectiva e aplicável de recomendações. Pode-se segmentar estas competências em dois tipos principais (Marler e Boudreau, 2016):

- Competências técnicas: identifica-se pelo processamento e tratamento de dados, posteriormente pela análise dos mesmos e, por fim, pela modelação estatística;

- Competências de negócio e comunicação: destaca-se pelo conhecimento do negócio, bem como das práticas e modelos de gestão de pessoas. A principal característica desta competência baseia-se na comunicação, como já foi referido anteriormente, de forma a que seja possível divulgar os resultados obtidos como se fosse uma história.

As competências técnicas e as de negócio e comunicação visam obter a melhor ligação tecnológica entre a gestão de pessoas.

A eficiência do processo variará proporcionalmente à qualidade e contributo dos *insights* produzidos na resposta aos temas críticos, no ponto de vista do negócio. As perguntas de partida revestem-se, assim, de uma importância fulcral para a análise e interpretação dos dados recolhidos. Identificado o problema, o processo de abordagem deve ser interpretado como se de uma cadeia de valor se tratasse. Escalar a cadeia desde a pergunta de partida à acção informada, requer uma abordagem criteriosa que permite compreender, medir, analisar e gerar medidas accionáveis.

A título de exemplo, supõe-se que a pergunta de partida resume-se a averiguar a existência de baixa produtividade na empresa. De seguida, procede-se à recolha de dados: as áreas *core* de negócio apresentam resultados de desempenho abaixo das demais áreas e tem-se sentido a dificuldade em contratar competências críticas para o desempenho das funções chave. Posteriormente, procede-se à análise dos dados recolhidos na etapa anterior. É necessário averiguar se existe oferta de talento no mercado, se a proposta de valor é atractiva e, consequentemente, analisar se existe eficácia no processo de recrutamento. Consequentemente, é fundamental contextualizar a estrutura do negócio da empresa. Como tal, é necessário analisar o aumento da procura de talento em contrapartida com as poucas as ofertas académicas em Portugal. É preciso interpretar as análises efectuadas no contexto da organização, por exemplo, se existe escassez de talento no mercado, aliada a condições laborais menos competitivas, como o prémio de desempenho ser abaixo da mediana do mercado. Posto isto, é necessário gerar acções, tendo como exemplo, a revisão da proposta de valor e melhoria da agilidade nos processos de recrutamento (Empresa, 2017).

Para optimizar o processo de resposta é importante recolher e analisar os temas mais críticos, em torno, a título de exemplo, de quatro categorias chave de desempenho, que orientam a segmentação dos indicadores mais importantes para a optimização das práticas de gestão de pessoas. Tendo como exemplo

as seguintes categorias: eficiência operacional, desenvolvimento de talento, *performance* financeira e serviço de RH. A eficiência operacional mede a competência e a melhoria contínua dos processos de RH e, o desenvolvimento de talento quantifica a robustez das políticas de gestão de talento. Relativamente à *performance* financeira e ao serviço de RH pode-se dizer que, respectivamente, quantificam o custo dos processos e modelos de RH, e a satisfação dos *stakeholders*, que consiste no grupo de interesse, ou seja, os colaboradores da empresa.

Neste processo, a limpeza dos dados afigura-se crítica para garantir a qualidade e a consistência destes, por forma a assegurar indicadores sólidos e confiáveis (Empresa, 2017).

A etapa da tecnologia baseia-se na definição de uma estratégia tecnológica para suportar a evolução analítica desejada. Desde o desenho da estrutura de dados e da arquitectura de sistema, às tecnologias que melhor respondem às necessidades da estratégia de evolução da função de HRA. Dada a sensibilidade da informação a disponibilizar na plataforma analítica, é necessário definir uma matriz de acessos, que permite que os principais *stakeholders* acedam a toda a informação necessária, de forma a optimizar a tomada de decisão, salvaguardando em simultâneo dados confidenciais.


Figura 1.2: Perfis de utilizadores da plataforma.

Desta forma, pretende-se que a matriz de acessos à plataforma seja consoante o tipo de acessos representados na figura 1.2.

A gestão da mudança resume-se à implementação da função de PA. O sucesso dos *insights* dependerá obrigatoriamente do envolvimento do negócio com os mesmos. A gestão da mudança assenta em 4 pilares: comunicação, auditoria, capacitação e apoio operacional. Destes pilares destaca-se a comunicação, uma vez que, será determinante para a efectivação das acções propostas (Empresa, 2017).

## 1.2 *Flight Risk*

Actualmente, as empresas confrontam-se com o facto de um colaborador com bastante potencial na empresa e, que por sua vez, executa as suas funções com um elevado desempenho, manifestar o interesse em rescindir o contracto com a entidade patronal. Consequentemente, a equipa da empresa que se destina ao recrutamento de talento vê-se, de certa forma, obrigada a procurar um potencial talento para preencher o cargo o mais rapidamente possível, seja na forma de recrutamento interno ou externo.

Este episódio reflecte uma situação que nenhuma empresa, seja de que dimensão for, quer enfrentar. Como tal, deve-se efectuar uma avaliação de risco de saída dos colaboradores, o que permite antecipar a rotatividade dos demais. Desta forma, o *Flight Risk* corresponde à previsão de saída voluntária de um

colaborador de uma empresa. Não se deve subestimar este indicador de negócio, uma vez que, se está perante a um mercado onde a competição é cada vez mais implacável e, consequentemente, as empresas têm que centrar o seu modelo de negócio no colaborador.

Em 2017, a economia do trabalho dos EUA favoreceu enormemente o empregado em detrimento do empregador, de acordo com os dados do Bureau of Labor Statistics. O número de posições de empregos não-agrícolas é superior a 5.5 milhões, correspondendo ao número mais elevado já registado. Como tal, a taxa de desemprego registada é de 4,7%. Pode-se concluir que, a partir de 2017 os trabalhadores que antes se contentavam em trabalhar em situações pouco satisfatórias, do ponto de vista salarial, têm a oportunidade de explorar outras opções, devido à elevada oferta de empregos (Westfall, 2017).

Posto isto, devido à elevada rotatividade dos colaboradores no mercado empresarial, os gestores de negócio de cada organização precisam de cobrir o custo associado à nova contratação. Segundo o estudo efectuado pelo Center for American Progress, o valor associado ao custo de uma nova contratação pode variar consoante as responsabilidades do cargo do colaborador, ou seja, para empregos de carácter administrativo e compostos por poucas responsabilidades a nível do negócio, o custo poderá corresponder até 16% do salário anual. Enquanto, em relação a posições de cargos de chefia e, que por sua vez, exigem uma maior componente analítica e estratégica, o custo associado poderá atingir os 213% do salário anual.

De acordo com Helen Poitevin e Alexander Linden, efectuar uma avaliação adequada aos riscos associados à rotatividade dos colaboradores, pode contribuir para a diminuição dos custos associados às novas contratações. Consequentemente, é possível identificar qual é o padrão do comportamento dos colaboradores que constituem um elevado risco de rescindir o contrato de forma voluntária. O estudo do risco de saída dos colaboradores pode implicar uma redução significativa das despesas da empresa e, consequentemente, evitar a perda de produtividade da organização (Siegel, 2013).

Alcançar os indicadores de risco que potencializam o risco de saída dos colaboradores tem sido um enorme desafio encarado pelas empresas, por exemplo, a Hewlett-Packard (HP).

A empresa HP alcançou um novo poder empresarial ao prever o comportamento dos colaboradores. Esta prática permitiu à empresa identificar mais de 330 mil funcionários com características de elevado risco de saída. Com a utilização desta análise preditiva, as organizações ganham poder ao construir *insights*, de forma a antecipar a saída dos colaboradores. Ao prever quais são os colaboradores potenciais a rescindir o contrato, a HP pode concentrar os recursos e esforços necessários em retê-los, reduzindo assim o alto custo associado às novas contratações.

Desta forma, a HP acredita que a capacidade de prever as saídas permite diminuir a taxa de rotatividade da força de trabalho. Permitindo, assim, a que a equipa de gestão de talento forneça suporte especializado para a gestão de pessoas. As taxas de rotatividade para uma amostra de, aproximadamente 300 colaboradores, eram cerca de 20%. Após a utilização da análise preditiva, a HP conseguiu uma redução de cinco pontos percentuais, com vista a uma redução contínua.

Além deste sucesso inicial, a capacidade de previsão do risco de saída da HP, permitiu à empresa uma poupança estimada de 300 milhões de unidades monetárias em relação à substituição de pessoal e perda de produtividade (Siegel, 2013).

A análise efectuada pela empresa permitiu concluir que o risco associado à saída dos colaboradores pode depender de factores definidos pela empresa, nomeadamente a retribuição salarial e a classificação da avaliação de desempenho. Os colaboradores com salários mais elevados e sujeitos às maiores promoções salariais são menos propensos a sair. Este novo poder da HP traduz-se numa ameaça para os colaboradores, uma vez que se baseia em dados pessoais e financeiros dos mesmos.

Para além disto, os colaboradores que estão inseridos na modelação desta análise questionam-se sobre os possíveis erros de previsão e, consequentemente, a caracterização errada. Todavia, a empresa

defende que este projecto não tem o intuito de penalizar os funcionários, sendo que o principal objectivo é minimizar as saídas voluntárias dos mesmos. Apesar das preocupações, a análise preditiva não invade a privacidade de cada colaborador (Siegel, 2013).

A Google é também um exemplo de uma empresa que recorreu ao uso da HRA, para determinar quais são os potenciais colaboradores a saírem da empresa de forma voluntária. Uma das conclusões obtidas do estudo efectuado é que os colaboradores que não recebem uma promoção num prazo de quatros anos, são classificados como os que têm maior probabilidade de rescindir o contracto.

# 2. Metodologias

O *data science* é a disciplina de processamento e análise de dados com a finalidade de obter conhecimento valioso. O termo *data science* foi criado na década de 1960, no entanto, só ganhou forma quando a tecnologia se tornou suficientemente madura.

Em diversas áreas de negócio, como por exemplo nos recursos humanos, na área da saúde e na investigação, conseguiu-se através do *data-driven* e das predições obter novos *insights*. Por exemplo, a utilização deste tipo de metodologia pela Google permitiu-lhe melhorar a relevância dos resultados dos seus motores de busca e gerir as campanhas publicitárias.

O *data mining* é a ciência, a arte e a tecnologia da exploração de grandes e complexos grupos de dados, com o intuito de descobrir padrões e gerar *insights*. Os cientistas procuraram continuamente técnicas aprimoradas para tornar o processo mais eficiente, económico e preciso. Um dos principais objectivos do *data mining* é permitir fazer previsões sobre certos fenómenos (Rokach e Maimon, 2015).

Inicialmente será feita uma introdução do modelo linear generalizado e do seu surgimento. De forma a enquadrar a metodologia utilizada neste projecto, é apresentado um caso particular dos modelos lineares generalizados, nomeadamente a regressão logística. De seguida, é apresentado o método a utilizar na selecção das variáveis a incluir no modelo e, por sua vez, como se deve interpretar os coeficientes do mesmo, de forma a que, seja possível actuar na empresa em estudo, consoante os *insights* obtidos. Posto isto, é necessário avaliar a capacidade discriminatória do modelo, através da análise de resíduos e de diversas métricas de avaliação de desempenho.

De forma a complementar este trabalho e comparar os resultados obtidos através da metodologia mencionada anteriormente, é apresentada outra abordagem, nomeadamente as árvores de decisão. Para esta metodologia, apresenta-se como é realizada a criação de uma árvore de decisão e, por sua vez, os critérios utilizados para avaliar o ajustamento do modelo.

## 2.1 O Modelo Linear Generalizado

Um modelo é uma abstracção da realidade uma vez que fornece uma aproximação de um qualquer fenómeno relativamente mais complexo (Myres et al., 2010). Como tal, o objectivo é encontrar o melhor ajustamento e mais parcimonioso, que permita estabelecer relações entre a variável resposta - variável dependente - e as variáveis explicativas - variáveis independentes.

O modelo linear normal, "criado" no início do século XIX por Legendre e Gauss, dominou a modelação estatística até meados do século XX, embora vários modelos não lineares ou não normais tenham entretanto sido desenvolvidos para fazer face a situações que não eram adequadamente explicadas pelo modelo linear normal. São exemplo disso, tal como referem McCullagh and Nelder (1989) e Lindsey (1997), o modelo *complementar log-log* para ensaios de diluição (Fisher, 1922), os modelos *probit* (Bliss, 1935) e *logit* (Berkson, 1944; Dyke and Patterson, 1952; Rasch, 1960) para proporções, os modelos *log-lineares* para dados de contagens (Birch, 1963), e os modelos de regressão para análise de

sobrevivência (Feigl and Zelen, 1965; Zippin and Armitage, 1966; Glasser, 1967).

Todos os modelos anteriormente descritos apresentam uma estrutura de regressão linear e têm em comum o facto da variável resposta seguir uma distribuição dentro de uma família de distribuições com propriedades muito específicas: a *família exponencial*.

Os modelos lineares generalizados introduzidos por Nelder e Wedderburn (1972) correspondem a uma síntese destes e de outros modelos, vindo assim unificar, tanto do ponto de vista teórico como conceptual, a teoria da modelação estatística até então desenvolvida. São casos particulares dos modelos lineares generalizados, os seguintes modelos (Turkman e Silva, 2000):

- Modelo de regressão linear clássico;
- Modelos de análise de variância e covariância;
- Modelo de regressão logística;
- Modelo de regressão de Poisson;
- Modelos *log-lineares* para tabelas de contingência multidimensionais;
- Modelo *probit* para estudos de proporções, etc.

Os modelos poderão ser classificados como determinísticos ou probabilísticos. Quando se está perante um modelo determinístico, os valores das variáveis explicativas são perfeitamente controlados pelo experimentador (Alpuim, 2018). No caso dos modelos probabilísticos, a variável resposta exibe variabilidade, uma vez que o modelo contém elementos aleatórios ou sofre o impacto de forças aleatórias. A classe de modelos probabilísticos mais importante é a classe dos modelos lineares:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon, \tag{2.1}$$

onde $Y$ corresponde à variável resposta, $X_1, X_2, ..., X_k$ é o conjunto de variáveis explicativas, $\beta_0, \beta_1, ..., \beta_k$ é o conjunto de parâmetros de regressão desconhecidos e $\varepsilon$ é o erro aleatório. À equação (2.1) dá-se o nome de modelo linear. Um dos principais pressupostos do modelo linear é que o valor médio de $\varepsilon$ é zero. Assim sendo, o valor médio da variável resposta é dado pela equação (2.2).

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{2.2}$$

O modelo linear generalizado é, tal como o nome indica, uma generalização do modelo linear para casos em que a variável resposta segue uma distribuição pertencente à família exponencial. A família exponencial inclui distribuições discretas e contínuas tais como a distribuição Normal, a distribuição Binomial, a distribuição Poisson, a distribuição Geométrica, a distribuição Binomial Negativa, a distribuição Exponencial e a distribuição Gama.

Diz-se que uma variável aleatória (v.a.) Y tem distribuição pertencente à família exponencial biparamétrica se a sua função de densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) se puder escrever na forma

$$f(y|\theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\} \tag{2.3}$$

onde $\theta$ é a forma canónica do parâmetro de localização, $\phi$ é um parâmetro de dispersão suposto, em geral, conhecido e $a(\cdot), b(\cdot)$ e $c(\cdot)$ são funções reais conhecidas. Admite-se, ainda, que a função $b(\cdot)$ é diferenciável e que o suporte da distribuição não depende dos parâmetros (Turkman e Silva, 2000).

Se Y for uma v.a. com distribuição pertencente à família exponencial, tal como definida em (2.3), tem-se que:

$$E[Y] = \mu = b'(\theta) \tag{2.4}$$

$$Var[Y] = a(\theta)b''(\theta), \tag{2.5}$$

onde $b'(\theta) = \frac{\partial b(\theta)}{\partial \theta}$ e $b''(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta^2}$.

Como exemplo, considere-se $\mathcal{T}$ uma v.a. com distribuição Binomial com parâmetros $n$ e $p$, $\mathcal{T} \sim Bin(n,p)$ e com f.m.p dada por:

$$\begin{aligned} P(\mathcal{T}=t) &= \binom{n}{t} p^t\ (1-p)^{n-t} \\ &= \exp\left\{ \ln\binom{n}{t} + t\ \ln(p) + (n-t)\ \ln(1-p) \right\} \\ &= \exp\left\{ \ln\binom{n}{t} + t\ \ln\left(\frac{p}{1-p}\right) + n\ \ln(1-p) \right\} \\ &= \exp\left\{ t\ \theta - n\ \ln(1+e^{\theta}) + \ln\binom{n}{t} \right\}, \end{aligned} \tag{2.6}$$

com $\theta = ln(\frac{p}{1-p})$ e $t = 0,1,2,3,\ldots,n$.

Assim, esta f.m.p. é da forma (2.6) com:

- $\theta = ln(\frac{p}{1-p})$, que corresponde à função *logit*;
- $b(\theta) = n\ ln(1+e^{\theta})$;
- $c(y,\phi) = ln\binom{n}{t}$;
- $a(\phi) = 1$;
- $b'(\theta) = n\ \frac{e^{\theta}}{1+e^{\theta}} = n\ p$;
- $a(\theta) \cdot b''(\theta) = 1\ n\ \frac{e^{\theta}}{(1+e^{\theta})^2} = n\ p\ (1-p)$

A generalização do modelo linear é obtida através da extensão das hipóteses subjacentes ao modelo linear. Esta extensão é feita em duas direcções: a relaxação da condição de que a variável resposta Y siga uma distribuição Normal, podendo seguir qualquer distribuição pertencente à família exponencial, a função que relaciona o valor esperado e o vector de covariáveis (chamada função de ligação e representada por $g(\cdot)$) pode ser qualquer função diferenciável. Formalmente, o modelo linear generalizado pode ser definido como na equação (2.7) (Myres et al., 2010).

$$g[\mu] = q[E(Y)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{2.7}$$

Qualquer construção de um modelo linear generalizado envolve 3 partes:

- uma distribuição para a variável resposta Y;

- um preditor linear que envolva as variáveis regressoras

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k; \tag{2.8}$$

- uma função de ligação invertível $g(\cdot)$ que relaciona o valor médio da variável resposta, $\mu \equiv E[Y]$, com o preditor linear.

Quando a função de ligação entre o valor médio da variável resposta e o preditor linear é a função *logit* (correspondendo a uma variável resposta com distribuição Binomial $(1, p)$), o modelo linear generalizado toma o nome particular de modelo de regressão logística, ou modelo *logit*.

### 2.1.1 Regressão Logística

Os modelos de regressão tornaram-se numa componente essencial de qualquer análise de dados relacionada com uma possível relação entre uma variável resposta e uma ou mais variáveis explicativas. Em relação ao estado da arte de HRA, na maioria dos casos, a variável dependente é discreta, assumindo dois ou mais valores possíveis.

Um dos casos particulares dos modelos lineares generalizados é o modelo de Regressão Logística, usado para predizer o resultado de uma variável dependente categórica (binária) baseada numa ou mais variáveis preditoras. Um modelo de regressão logística tanto pode ser binomial como multinomial. No caso da Regressão Logística binomial, a variável resposta assume valores distintos, normalmente 0 e 1, sendo 1 a codificação atribuída ao sucesso de um determinado acontecimento, cuja ocorrência se pretende predizer (Scott et al., 2013). O modelo de Regressão Logística é o modelo mais utilizado para análises de dados, comumente utilizado no estado da arte de HRA.

A inferência estatística em modelos de regressão linear assenta no pressuposto de que os termos de erro são variáveis independentes e identicamente distribuídas com distribuição normal. Consequentemente, as observações da variável dependente são também independentes e com distribuição normal. Os métodos de estimação existentes, os testes e intervalos de confiança, são bastante robustos relativamente a este pressuposto. Como tal, mesmo que os termos de erro se afastem da distribuição normal, para amostras de dimensão grande, é possível aplicar métodos de inferência estatística, não tendo o risco de cometer grandes erros. Esta característica dos modelos lineares deve-se ao Teorema do Limite Central e suas generalizações para variáveis não identicamente distribuídas, e da Lei Fraca dos Grandes Números (Alpuim, 2018).

Na Regressão Logística (Dobson e Barnett, 2008), a variável dependente, Y, é uma variável dicotómica que pode ser descrita na forma

$$Y = \begin{cases} 0, & \text{se o } \textit{outcome} \text{ é um insucesso} \\ 1, & \text{se o } \textit{outcome} \text{ é um sucesso} \end{cases}, \tag{2.9}$$

sendo uma variável com distribuição binomial de parâmetros 1 e $\pi$, $Y \sim Bin(1, \pi)$, por outras palavras, segue uma distribuição Bernoulli, $Y \sim Ber(\pi)$, cuja função massa de probabilidade é dada por

$$f(y|\pi) = \pi^y \, (1-\pi)^{1-y}, \; y = 0, 1, \tag{2.10}$$

com probabilidades $P(Y = 1) = \pi$ e $P(Y = 0) = 1 - \pi$.

A Regressão Logística é utilizada para predizer os *odds* da resposta $Y = 1$ baseado nas variáveis preditivas. Os *odds* de $Y = 1$ são definidos como a probabilidade da variável resposta tomar o valor 1

dividida pela probabilidade de tomar o valor 0. Formalmente,

$$Odds = \frac{P(Y=1)}{P(Y=0} = \frac{\pi}{1-\pi} \tag{2.11}$$

De acordo com Scott et al., 2013, no caso de um modelo de Regressão Logística simples, em qualquer regressão a quantidade chave é o valor médio da variável resposta, $Y$, dado o valor da variável independente $X$, isto é, o valor médio condicional, e é expressa na forma

$$\pi(x) = E[Y|X=x] \tag{2.12}$$

e, tem-se:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{2.13}$$

Como referido na Secção 2.1, e segundo Scott et al., 2013, a função de ligação utilizada na função logística e, por sua vez, uma das transformações fulcrais é designada por transformação *logit*,

$$g(x) = ln\Big(\frac{\pi(x)}{1-\pi(x)}\Big) = \beta_0 + \beta_1 x \tag{2.14}$$

Consequentemente,

$$\frac{\pi(x)}{1-\pi(x)} = \frac{\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}}{1-\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}} = e^{\beta_0+\beta_1 x} \tag{2.15}$$

A *logit* pode assumir qualquer valor pertencente ao intervalo $(-\infty;+\infty)$, uma vez que $\pi(x)$ só pode tomar valores entre [0,1].

Desta forma, é também possível aplicar o modelo de Regressão Logística para descrever a probabilidade de um determinado acontecimento como uma função logística multivariada de um conjunto de variáveis independentes, $X_1, X_2, ..., X_p$. Assim, a função logística é escrita na forma

$$\pi(\underset{\sim}{x}) = \frac{e^{\beta_0+\beta_1 x+\beta_2 x_2+...+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x+\beta_2 x_2+...+\beta_p x_p}} \tag{2.16}$$

A função logística multivariada permite incluir transformações de uma mesma variável $x$ como, por exemplo, potências de $x$, o que lhe dá maior flexibilidade e possibilidade de adquirir uma maior variedade de formas (Alpuim, 2018).

#### 2.1.1.1 Ajustamento do Modelo

Considere-se uma amostra composta por $n$ observações independentes do par $\big(y_i, \underset{\sim}{x_i}\big)$, $i = 1,2,...,n$, onde $y_i$ corresponde ao valor da variável dicotómica e $\underset{\sim}{x_i}$ é a i-ésima observação do vector de variáveis independentes. Para se ajustar um modelo de regressão logística múltiplo é necessário estimar os parâmetros desconhecidos, $\beta_0, \beta_1, ..., \beta_p$.

O método de estimação dos parâmetros utilizado na regressão linear é o Método dos Mínimos Quadrados, que consiste na minimização da soma dos quadrados dos resíduos, isto é, na minimização da soma dos quadrados das diferenças entre os valores observados $Y$ e os valores estimados do modelo

(RSS - *Residual Sum of Squares*).

$$RSS(\beta_0, \beta_1, ..., \beta_p) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.17}$$

Sob as condições usuais de um modelo de regressão linear, o Método dos Mínimos Quadrados produz estimadores com um número de propriedades estatísticas desejáveis. No entanto, quando este método é aplicado a um modelo cuja variável resposta é dicotómica, os estimadores não apresentam as mesmas propriedades.

Quando os erros aleatórios seguem uma distribuição Normal, o método que se traduz na função de mínimos quadrados do modelo de regressão linear, particularmente na regressão logística, designa-se por Método da Máxima Verosimilhança.

Este método produz estimadores que maximizam a probabilidade de obter o conjunto de dados observados. Para que seja possível estimar os parâmetros desconhecidos, é necessário encontrar a função de verosimilhança, $L(\beta)$. Esta função representa a distribuição conjunta dos dados observados.

Assumindo uma amostra de dimensão $n$, em que as observações são independentes, a função de verosimilhança é dada por

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \tag{2.18}$$

De forma a simplificar o cálculo da função $L(\beta)$, sabe-se que maximizar $L(\beta)$ é equivalente a maximizar o seu logaritmo. Assim, a log-verosimilhança é escrita na forma

$$\ell(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n} \left\{ y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)] \right\} \tag{2.19}$$

Tendo em consideração as derivadas da função log-verosimilhança em ordem aos parâmetros $\beta_0$ e $\beta_k$, é possível obter os estimadores de máxima verosimilhança para o conjunto de $\beta$ desconhecidos como solução do sistema de $(p+1)$ equações normais

$$\frac{\partial \ln L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \tag{2.20}$$

$$\frac{\partial \ln L(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} x_{ik} [y_i - \pi(x_i)] = 0, \tag{2.21}$$

em que $k = 1, ..., p$, ou, alternativamente, em notação matricial

$$X'(Y - \Pi) = 0. \tag{2.22}$$

Os estimadores de máxima verosimilhança de $\beta$, que corresponde ao vector de parâmetros desconhecidos, são obtidos como solução das equações de verosimilhança. A solução não corresponde necessariamente a um máximo global da função $\ell(\beta)$. Contudo, em muitos modelos a função log-verosimilhança é côncava, de modo que o máximo local e global coincidem. Para funções estritamente côncavas, os estimadores de máxima verosimilhança são únicos, quando existem. No entanto, relativamente ao problema da existência e unicidade destes estimadores não existem soluções para esta questão, uma vez que nem todos os modelos têm propriedades comuns. Partindo do princípio que existe solução e esta é única, subsiste ainda o problema com o cálculo das estimativas de máxima verosimilhança, uma vez

que as equações de verosimilhança não têm, em geral, solução analítica, e, portanto, implica o recurso a métodos numéricos (Turkman e Silva, 2000).

Assim, utiliza-se, por exemplo, o algoritmo de Newton-Raphson para obter os estimadores de máxima verosimilhança para $\beta$. Este algoritmo iterativo inicia-se com uma solução inicial, alterando-a ligeiramente de forma a perceber se pode ser melhorada e repete-se este processo sucessivamente até que o melhoramento atinja uma certa precisão. Quando tal acontece diz-se que o processo iterativo converge.

No entanto, existem casos em que não é possível atingir a convergência. Neste casos, uma vez que o processo iterativo não foi capaz de fornecer uma solução, os coeficientes do modelo de regressão não são significativos (Menard, 1995). Isto pode acontecer por diversas razões, como por exemplo devido à existência de multicolinearidade.

O conceito de multicolinearidade refere-se a uma correlação inaceitavelmente alta entre preditores, isto é, quando existe uma relação de (quase) dependência linear entre os vectores de valores observados das variáveis. Uma perfeita multicolinearidade significa que pelo menos uma variável explicativa é uma combinação linear perfeita das outras. Se cada variável independente fosse analisada como variável dependente num modelo com todas as restantes variáveis explicativas, a multicolinearidade perfeita resultaria num $R^2$ de 1, para pelo menos uma das variáveis, o que se traduz na impossibilidade de obter uma estimativa única dos coeficientes de regressão. Quando a multicolinearidade aumenta, os coeficientes mantêm-se centrados, no entanto os erros padrões aumentam e a verosimilhança do modelo diminui (Menard, 1995).

Outra forma de detectar a existência de multicolinearidade entre os preditores é calcular o *Variance Inflation Factor* (VIF). Esta quantidade para cada um dos preditores é escrita na forma

$$VIF_k = \frac{1}{1-R_k^2}, \tag{2.23}$$

onde $R_k^2$ é o valor do Coeficiente de Determinação ($R^2$) obtido através da regressão do preditor $k$ nos restantes preditores. O Coeficiente de Determinação é uma medida de ajustamento que pode assumir qualquer valor pertencente ao intervalo $(0,1)$, indicando, em percentagem, a quantidade de variabilidade dos dados que é explicada pelo modelo de regressão ajustado (Dobson e Barnett, 2008). Ou seja, quanto maior for o valor do $R^2$ melhor é o ajustamento do modelo à amostra e, consequentemente, mais explicativo. Claramente, se $x_k$ é quase linearmente dependente de alguns dos outros preditores, então $R_k^2$ estará próximo de 1 e, consequentemente, o $VIF_k$ será grande (Montegomery et al., 2012).

Está convencionado que quando um valor de *VIF* é superior a 5, pode-se afirmar que se está perante uma situação de multicolinearidade elevada (apesar de haver alguns autores que tomam valor 10 como valor fronteira (Kutner et al., 2004)). Caso um preditor não esteja correlacionado com os restantes preditores, então o VIF é igual a 1.

Tal como num modelo de regressão linear, na qual é utilizada a soma de quadrados, é necessário avaliar a qualidade de ajustamento do modelo. No caso do modelo de regressão logística, a *Deviance* é análoga à soma de quadrados, correspondendo à medida dos desvios no ajuste de um modelo de regressão logística aos dados em causa. A este método também se dá o nome de teste de Razão de Verosimilhança (Scott et al., 2013).

O modelo saturado, que corresponde ao modelo completo por todas as variáveis independentes, é útil para julgar da qualidade de ajustamento de um determinado modelo em investigação, que passamos a designar por M, através da introdução de uma medida da distância dos valores ajustados $\hat{\mu}$ com esse modelo e dos correspondentes valores observados $y$. Essa medida de discrepância entre o modelo saturado e o modelo corrente, é baseada na estatística de razão de verosimilhanças de Wilks. A estatística de

Wilks ou estatística de razão de verosimilhanças é definida por

$$\Lambda = -2 \ln \frac{\max_{H_0} L(\beta)}{\max_{H_0 \cup H_1} L(\beta)} = -2\{\ell(\tilde{\beta}) - \ell(\hat{\beta})\} \tag{2.24}$$

O logaritmo da função de verosimilhança (*função log-verosimilhança*) de um modelo linear generalizado é dado por (Turkman e Silva, 2000)

$$ln\, L(\beta) = \ell(\beta) = \sum_{i=1}^{n} \frac{\omega_i [y_i q(\mu_i) - b(q(\mu_i))]}{\phi} + c(y_i, \phi, \omega_i) \tag{2.25}$$

em que se substitui $\theta_i$ por $q(\mu_i)$, para fazer salientar, na função *log-verosimilhança*, a relação funcional existente entre $\theta_i$ e $\mu_i$.

Como para o modelo saturado - designado por $\mathcal{S}$ - se tem $\hat{\mu}_i = y_i$, o máximo da *função log-verosimilhança* para este modelo é

$$\ell_S(\hat{\beta}_S) = \sum_{i=1}^{n} \frac{\omega_i [y_i q(y_i) - b(q(y_i))]}{\phi} + c(y_i, \phi, \omega_i) \tag{2.26}$$

Por outro lado, se se designar por $\hat{\mu}_i$ a estimativa de máxima verosimilhança de $\mu_i$, para $i = 1, ..., n$, o máximo da *função log-verosimilhança* para o modelo em investigação com $m$ parâmetros no desvio é

$$\ell_M(\hat{\beta}_M) = \sum_{i=1}^{n} \frac{\omega_i [y_i q(\hat{\mu}_i) - b(\hat{\mu}_i)]}{\phi} + c(y_i, \phi, \omega_i) \tag{2.27}$$

Os índices em $\hat{\beta}$ e $\ell$ correspondem ao modelo em relação ao qual são calculados. Ao comparar o modelo em investigação M com o modelo saturado S através da estatística de razão de verosimilhanças, obtém-se

$$\begin{aligned} D^*(y; \hat{\mu}) &= -2(\ell(\hat{\beta}_M) - \ell(\hat{\beta}_S)) \\ &= -2 \sum_i \frac{\omega_i}{\phi} \Big\{ \Big[ y_i q(\hat{\mu}_i) - b(q(\hat{\mu}_i)) \Big] - \Big[ y_i q(y_i) - b(q(y_i)) \Big] \Big\} \\ &= \frac{D(y; \hat{\mu})}{\phi} \end{aligned} \tag{2.28}$$

À $D^*(y; \hat{\mu})$ definida em (2.28) dá-se o nome de desvio reduzido; o numerador $D(y; \hat{\mu})$ designa-se por desvio para o modelo corrente (Turkman e Silva, 2000).

Assim, a *Deviance* é definida por:

$$\begin{aligned} D &= -2 \ln \Big[ \frac{\text{Função de máx. verosimilhança do modelo nulo}}{\text{Função de máx. verosimilhança do modelo saturado}} \Big] \\ &= -2 \sum_{i=1}^{n} \Big[ y_i \ln \Big( \frac{\hat{\pi}_i}{y_i} \Big) + (1 - y_i) \ln \Big( \frac{1 - \hat{\pi}_i}{1 - y_i} \Big) \Big] \end{aligned} \tag{2.29}$$

A hipótese a testar é então:

$$H_0 : \beta_1 = ... = \beta_q = \beta_0 = 0 \;\; vs. \;\; H_1 : \exists_{j=1,...,p} : \beta_j \neq 0$$

A estatística de teste definida em (2.30), segue uma distribuição assintótica $\chi^2_p$ e, é dada por (Scott

et al., 2013):

$$\begin{aligned} G &= D \text{ (modelo sem as } p \text{ variáveis) - } D \text{ (modelo com as } p \text{ variáveis)} \\ G &= -2\ln\left[\frac{\text{Função de máx. verosimilhança do modelo sem as } p \text{ variáveis}}{\text{Função de máx. verosimilhança do modelo com as } p \text{ variáveis}}\right] \end{aligned} \tag{2.30}$$

É fácil de verificar que a *Deviance* é sempre maior ou igual a zero, e decresce à medida que covariáveis vão sendo adicionadas ao modelo nulo, tomando obviamente o valor zero para o modelo saturado.

Uma outra propriedade importante da *Deviance* é a aditividade para modelos encaixados. Com efeito, supõe-se que se está perante dois modelos intermédios $M_1$ e $M_2$, sendo que $M_2$ é encaixado em $M_1$, ou seja, são modelos do mesmo tipo, mas o modelo $M_2$ contém menos parâmetros na *Deviance* que o modelo $M_1$. Se se designar por $D(y;\ \hat{\mu}_j)$ a *Deviance* do modelo $M_j, j = 1,2$, então a estatística da razão de verosimilhanças para comparar estes dois modelos resume-se a

$$-2\ (\ell_{M_2}(\hat{\beta}_2)\ -\ \ell_{M_1}(\hat{\beta}_1)) = \frac{D\ (y;\ \hat{\mu}_2)\ -\ D\ (y;\ \hat{\mu}_1)}{\phi} \tag{2.31}$$

Então, sob a hipótese do modelo $M_1$ ser verdadeiro, tem-se

$$\frac{D\ (y;\ \hat{\mu}_2)\ -\ D\ (y;\ \hat{\mu}_1)}{\phi} \overset{a}{\sim} \chi^2_{p_1-p_2}, \tag{2.32}$$

onde $p_j$ representa a dimensão do vector $\beta$ para o modelo $M_j$, $j = 1,2$.

Outra forma de avaliar o ajustamento do modelo actual é ao utilizar o Critério de Informação de *Akaike/Akaike Information Criterion* (AIC), o qual é baseado na função log-verosimilhança, penalizando esse valor com o número de covariáveis do modelo. Um valor baixo para o AIC é considerado como representativo de um melhor ajustamento e na selecção de modelos deve-se ter como objectivo a minimização do valor de AIC (Turkman e Silva, 2000).

A medida AIC é uma ferramenta para a selecção de modelos. Perante um conjunto de dados e vários modelos candidatos, estes podem ser ordenados de acordo com o AIC, considerando-se o melhor modelo aquele que apresentar menor valor de AIC. Isto permite dizer que um modelo é preferível a outro mas não é possível estabelecer um valor para o AIC acima do qual um modelo deva ser "rejeitado" (Bermudez, 2019). A medida AIC é definida por

$$\text{AIC} = -2 \times [\ln(L) - k], \tag{2.33}$$

onde $k$ é o número de parâmetros do modelo e $L$ é o valor da verosimilhança do modelo ajustado.

Quanto maior for o número de variáveis consideradas no modelo e, consequentemente mais parâmetros, maior será o valor da verosimilhança, pelo que $ln(L)$ cresce com a complexidade do modelo. Por outro lado, porque um modelo mais complexo acarreta maiores custos (a todos os níveis), a introdução de variáveis no modelo é penalizada. No entanto, esta medida não fornece qualquer informação sobre a significância dos modelos (Bermudez, 2019).

#### 2.1.1.2 Método de Selecção de Variáveis

Na presença de um modelo múltiplo, à partida, todas as variáveis independentes serão consideradas como relevantes para a construção do modelo. Um modelo composto por um maior número de variáveis

consegue uma melhor explicação da variável dependente, no entanto, esse modelo não será, necessariamente, o melhor sob o ponto de vista de predição. Por outro lado, um aspecto de elevada importância é o da interpretabilidade do modelo, que fica simplificada se este não envolver um número demasiado elevado de variáveis.

Existem vários métodos que podem ser usados na procura do "melhor" modelo. Tendo pontos de partida diferentes, estes métodos não conduzem todos ao mesmo resultado nem tampouco reúnem consenso relativamente a qual apresenta maiores vantagens.

Considerando uma situação em que existem $m$ covariáveis, uma possibilidade seria ajustar:

- um modelo que contenha as $m$ covariáveis;
- os $\frac{m(m-1)}{2}$ modelos contendo todas as combinações de $m-1$ das $m$ variáveis;
- os $\binom{m}{k}$ modelos contendo todas as combinações de $k$ das $m$ variáveis, $k = m-2, ..., 1$;
- e para terminar, ajustar o modelo sem variáveis regressoras, ou seja, $E(Y) = \beta_0$.

Após o ajustamento de $\sum_{k=0}^{m} \binom{m}{k} = 2^m$, é possível escolher aquele modelo que produzisse menor erro quadrático médio ou, de forma equivalente, maior coeficiente de determinação ajustado ($R^2$) ou menor estimativa para o erro padrão, caso o objectivo do estudo fosse a predição.

A utilização desta metodologia é, obviamente, desaconselhada mesmo para problemas envolvendo um número relativamente reduzido de covariáveis dado o número de equações de regressão a estimar, para além de outras questões relacionadas com o critério de classificação do "melhor" modelo (Bermudez, 2019).

Relativamente aos processos utilizados para seleccionar o melhor modelo, o procedimento mais comum é o método *stepwise*. Este procedimento envolve inclusão e eliminação de variáveis que, de acordo com algum critério (usualmente o AIC), e partindo do modelo saturado (direcção *backward*) ou partindo do modelo nulo (direcção *forward*), escolhe o melhor modelo. Também é possível aplicar o método *stepwise* com a direção *both* que analisa as duas direcções (*backward e forward*) em simultâneo. À medida que é incluída/retirada uma variável do modelo, todas as variáveis são analisadas com o objectivo de determinar se deverá ser eliminada/adicionada no modelo naquele passo.

#### 2.1.1.3 Coeficientes do Modelo

Após o ajustamento do modelo, é usual examinar a contribuição de cada um dos preditores individuais para o ajustamento global. Para tal, é necessário examinar os coeficientes de regressão. Na regressão logística os coeficientes representam a mudança do *logit* por cada unidade de mudança no preditor. No caso de todos os preditores serem variáveis binárias, o coeficiente de cada preditor representa a mudança no *logit* caso o preditor tome o valor 1. No entanto, e uma vez que a interpretação do *logit* não é imediata, é usual focar-se no efeito que um preditor tem no *Odds Ratio* (OR). O *Odds Ratio* é usado como uma medida que descreve a força da associação (ou da não independência) entre dois conjuntos de observações de variáveis binárias.

A chance de um indivíduo classificado como positivo, $Y = 1$, face a um indivíduo não positivo é $\frac{\pi(1)}{[1-\pi(1)]}$. Da mesma forma, a chance de um indivíduo classificado como negativo, $Y = 0$, é $\frac{\pi(0)}{[1-\pi(0)]}$. O OR é a razão entre as probabilidades para $Y = 1$ e as probabilidades para $Y = 0$, e é dada pela equação:

$$\text{OR} = \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}} \tag{2.34}$$

Substituindo as expressões para as probabilidades do modelo de regressão logística tem-se:

$$\text{OR} = \frac{\dfrac{\dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}}{\dfrac{1}{1+e^{\beta_0+\beta_1}}}}{\dfrac{\dfrac{e_0^{\beta}}{1+e_0^{\beta}}}{\dfrac{1}{1+e_0^{\beta}}}} = \frac{e^{\beta_0+\beta_1}}{e_0^{\beta}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1} \tag{2.35}$$

Assim, para um modelo de regressão logística com uma variável independente dicotómica, a relação entre a razão de probabilidades e o coeficiente de regressão é (Scott et al., 2013)

$$\text{OR} = e^{\beta_1} \tag{2.36}$$

Ao contrário da regressão linear, na qual a importância de um coeficiente é avaliada utilizando um teste $t$, na regressão logística a significância de um preditor individual pode ser avaliada através do teste de razão de verosimilhanças ou da estatística de *Wald* (Turkman e Silva, 2000).

Considere-se que a hipótese nula estabelece que $C\beta = \xi$, onde $C$ é uma matriz $q$ x $p$ de característica completa $q$. Seja $\hat{\beta}$ o estimador de máxima verosimilhança de $\beta$, o qual tem uma distribuição assintótica $N_p\big(\beta, \mathcal{I}^{-1}(\beta)\big)$. Dado que o vetor $C\hat{\beta}$ é uma transformação linear de $\hat{\beta}$ então, pelas propriedades da distribuição normal multivariada,

$$C\hat{\beta} \overset{a}{\sim} N_q\big(C\beta, C\mathcal{I}^{-1}(\beta)C^T\big) \tag{2.37}$$

e, consequentemente, sob a hipótese nula, a estatística

$$\mathcal{W} = \big(C\hat{\beta} - \xi\big)^T \big[C\mathcal{I}^{-1}(\hat{\beta})C^T\big]^{-1}\big(C\hat{\beta} - \xi\big), \tag{2.38}$$

tem uma distribuição assintótica de um $\chi^2$ com $q$ graus de liberdade. À estatística $\mathcal{W}$ dá-se o nome de Estatística de *Wald*.

Assim, a hipótese nula é rejeitada a um nível de significância $\alpha$, se o valor observado da estatística de *Wald* for superior ao quantil de probabilidade $1-\alpha$ de um $\chi^2_q$.

Portanto, o objetivo é testar as seguintes hipóteses:

$$H_0 : \beta_j = 0 \;\; vs. \;\; H_1 : \beta_j \neq 0,$$

para algum $j$.

Então, utilizando (2.38) e designando por $\sigma_{jj}$ o j-ésimo elemento da diagonal de $\mathcal{I}^{-1}\big(\hat{\beta}\big)$, a estatística

de *Wald* resume-se a,

$$\mathcal{W} = (\hat{\beta}_j - \beta_j)^T \left[\sigma_{jj}\right]^{-1} (\hat{\beta}_j - \beta_j) \tag{2.39}$$

e, portanto, sob $H_0$,

$$\mathcal{W} = \frac{\hat{\beta}_j{}^2}{\sigma_{jj}} \overset{a}{\sim} \chi^2_1. \tag{2.40}$$

Quando o coeficiente da regressão tem um valor muito elevado, o desvio padrão do coeficiente também tende a ser elevado, o que aumenta a probabilidade de erro de tipo II (Menard, 1995). Quando se está perante uma amostra de pequena dimensão, existem outros métodos com um desempenho melhor que o da estatística de *Wald*, dado que esta é tendencialmente enviesada (Agresti, 2019).

#### 2.1.1.4 Diagnóstico do Modelo

Uma vez verificado o ajustamento do modelo e a significância de todos os coeficientes do mesmo, é necessário efectuar o diagnóstico do modelo de forma a garantir que as conclusões retiradas são as mais acertadas e com o menor erro possível e, por sua vez, identificar observações mal ajustadas.

Uma das formas de avaliar a qualidade de ajustamento do modelo é através da análise de resíduos. O uso de gráficos é o método mais utilizado para realizar esta análise. A existência de qualquer tendência do gráfico em análise pode indicar uma escolha errada da função de ligação, ou uma escolha errada da escala da covariável em questão (Turkman e Silva, 2000).

Define-se como resíduos de um modelo a diferença entre os valores observados e os valores ajustados:

$$r(i) = y(i) - \hat{y}(i) \tag{2.41}$$

Uma vez que se está perante a um modelo de regressão logística, como foi referido anteriormente, a variável de resposta é binária, isto é, apenas toma valores 0 ou 1 e, como $\hat{y}$ é o valor estimado do modelo, que varia no intervalo [0, 1], os resíduos do modelo apenas variarão entre [-1, 1]. Quando $r(i) > 0$ corresponde aos casos em que $y(i) = 1$ e, analogamente, caso $r(i) < 0$ corresponde aos casos em que $y(i) = 0$. Por sua vez, quando $r(i) = 0$ está-se perante a situação em que o ajustamento do modelo é perfeito, ou seja, $y(i) = \hat{y}(i)$.

Ao contrário dos modelos de regressão linear, nos quais existe uma componente aleatória à qual se pode associar o valor dos resíduos, no caso dos modelos lineares generalizados e, em particular, o modelo de regressão logística, essa componente não existe, fazendo, portanto, mais sentido considerar uma outra definição de resíduos (Portugal, 2013).

Uma alternativa consiste em considerar os resíduos de *Pearson* padronizados que são definidos como:

$$rp(i) = \frac{y(i) - \hat{y}(i)}{\sqrt{\hat{var}(Y(i))}}, \tag{2.42}$$

sendo $\sqrt{\hat{var}(Y(i))}$ uma estimativa do desvio padrão de $Y(i)$.

Os resíduos de *Pearson studentized* são dados por:

$$\frac{rp(i)}{\sqrt{1 - h_{ii}}}, \tag{2.43}$$

onde $h_{ii}$ representa o i-ésimo termo da diagonal da matriz de projecção generalizada $H$ (também chamada *hat matrix* uma vez que $\hat{y} = Hy$), que pode ser calculado através da expressão $h_{ii} = x(i)^T(X^TX)^{-1}x(i)$ (Antunes, 2009). A desvantagem dos resíduos de *Pearson* é que a sua distribuição é, geralmente, bastante assimétrica para os modelos não Normais, como é o caso do modelo de regressão logística (Portugal, 2013).

Tal como foi referido anteriormente, a análise de resíduos é feita, principalmente, com recurso a ferramentas gráficas. Entre os diferentes gráficos possíveis de se obter, destacam-se:

- *Scatterplot* dos resíduos - também conhecido como gráfico de dispersão dos resíduos. Permite verificar se os resíduos apresentam qualquer tipo de padrão, assim como se se encontram bem distribuídos no intervalo [-2, 2], sendo que, no mínimo, 95% dos resíduos se devem encontrar encontrar neste intervalo;

- Resíduos *vs*. valores ajustados - permite avaliar se a variância dos resíduos não é constante (existência de heterocedasticidade). Assume-se que os resíduos são independentes dos valores ajustados, querendo dizer que a correlação entre resíduos e valores preditos deve tomar valor 0;

- *Normal Q-Q plot* dos resíduos - permite avaliar se os resíduos seguem uma distribuição $Normal(0,1)$, por comparação aos quantis teóricos desta distribuição;

- Histograma dos resíduos - permite avaliar a simetria (ou assimetria) dos resíduos, podendo ajudar à detecção de padrões de resíduos.

A análise de resíduos é de alta importância uma vez que permite averiguar a existência de desvios sistemáticos do modelo. Para além desta análise, é também importante averiguar a existência de desvios isolados do modelo, isto é, averiguar a existência de uma ou mais observações mal ajustadas pelo modelo, designadas por observações discordantes (Turkman e Silva, 2000). É preciso ter em atenção que, dependendo da influência que estas observações têm aquando da estimação dos parâmetros, a existência de observações discordantes pode ser prejudicial para o modelo e, como tal, pode pôr em causa o desempenho do modelo em análise.

Assim, para averiguar se uma observação é, ou não, discordante e/ou influente é necessário reter dois conceitos importantes:

- *Leverage* - mede o efeito que a observação tem nos valores preditos, sendo um indicativo de quão influente uma observação é;

- Influência - uma observação é influente se, uma ligeira modificação, ou exclusão do modelo, produz alterações significativas nas estimativas dos parâmetros do modelo. A sua presença pode, por isso, originar um impacto indevido nas conclusões a retirar do modelo.

Assim, observações influentes não têm, necessariamente, resíduos elevados (Turkman e Silva, 2000).

Outra forma de encontrar observações influentes (com *leverage* elevada) e mal ajustadas (com resíduos grandes) é através da análise da distância de *Cook*. Esta medida de distância é um diagnóstico de exclusão, ou seja, mede a influência da i-ésima observação quando esta é removida da amostra (Montegomery et al., 2012).

A análise de *Cook* é dada por:

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}},\ i = 1,2,...,n \tag{2.44}$$

Desta forma, pode-se verificar que, além da constante $p$, $D_i$ é o produto do quadrado do i-ésimo resíduo *studentized* e $\frac{h_{ii}}{1-h_{ii}}$. Um resíduo *studentized* resulta do quociente entre o resíduo estimado e o seu desvio padrão. Existem diversos critérios para se considerar uma distância de *Cook* como elevada. Há autores que defendem que considerar $D_i > 1$ é suficiente para classificar uma distância como elevada e, por sua vez, considerar que a observação é influente (Montegomery et al., 2012). No entanto, existem vários autores que defendem que o valor desta distância deve ter em consideração a dimensão da amostra, $n$, ou seja a distância deve ser ponderada pelo número de observações usadas para fazer o ajuste do modelo, através do critério $D_i > 4/n$ (Bollen e Jackman, 1985). No âmbito deste projecto usar-se-á o primeiro critério.

Uma vez terminada a análise de resíduos é necessário medir a capacidade discriminatória do modelo. Partindo de um conjunto de observações distintas das utilizadas para ajustar o modelo, para as quais é conhecido o valor da variável resposta, faz-se a predição utilizando o modelo ajustado. Para cada uma das observações o modelo dará uma propensão (valor entre 0 e 1) que poderá ser interpretada como a probabilidade de que essa mesma observação apresente valor 1 para a variável resposta (Portugal, 2013).

Como os valores preditos pelo modelo são contínuos, ou seja, variam no intervalo [0, 1], é necessário definir um ponto de corte, *cut-off*, para ser possível classificar e contabilizar o número de predições positivas (quando o valor predito é superior ao *cut-off* e, por sua vez, a variável resposta toma valor 1) e negativas (analogamente, quando o valor predito é inferior ao *cut-off* e, como tal, a variável resposta toma valor 0).

Uma vez fixado o valor do *cut-off* é possível dividir as propensões em duas categorias distintas: abaixo do *cut-off*, (assume-se o valor 0) e acima do *cut-off* (assume-se valor 1).

Tabela 2.1: Matriz de Confusão.

| | **Observados** | |
|---|---|---|
| **Previstos** | 1 | 0 |
| 1 | VP | FP |
| 0 | FN | VN |

A matriz de confusão, representada na tabela 2.1, é usada como indicação das propriedades de uma regra de classificação. Esta matriz contém o número de elementos que foram classificados corretamente ou incorrectamente para cada cenário possível (Rokach e Maimon, 2015).

Como para este conjunto de observações é conhecido o valor real da variável resposta, é possível compará-lo com a categoria da predição feita pelo modelo. Assim, está-se perante quatro cenários possíveis:

- Verdadeiros Positivos (VP) - A categoria de predição é 1 e o valor observado da variável resposta é também 1;
- Falsos Positivos (FP) - A categoria de predição é 1 e o valor observado da variável resposta é 0;
- Verdadeiros Negativos (VN) - A categoria de predição é 0 e o valor observado da variável resposta é também 0;
- Falsos Negativos (FN) - A categoria de predição é 0 e o valor observado da variável resposta é 1.

Naturalmente, o que se pretende é que os números totais de falsos positivos e de falsos negativos sejam os menores possíveis. Estes valores podem ser alterados modificando o *cut-off*. Contudo, se, por

exemplo, se aumentar o *cut-off*, diminui o número de falsos positivos, mas aumenta o número de falsos negativos. Inversamente, se se diminuir o valor do *cut-off*, diminui também o número de falsos negativos mas aumenta o número de falsos positivos. Ou seja, por modificação do *cut-off*, não é possível diminuir o número de falsos positivos sem aumentar os falsos negativos, bem como inversamente (Alpuim, 2018).

Com o objectivo de avaliar a qualidade do modelo, baseado nos valores da tabela 2.1, é possível calcular as seguintes medidas (Portugal, 2013, Rokach e Maimon, 2015 e Alpuim, 2018):

- Sensibilidade - corresponde à taxa de verdadeiros positivos, ou seja, representa a capacidade do modelo predizer correctamente as observações que se encontram na categoria 1 da variável de interesse. De outra forma, representa a probabilidade de uma observação ser classificada como positiva dado que é, de facto, positiva. É dada por:

$$S = \frac{\text{Número de verdadeiros positivos}}{\text{Total de positivos}} = \frac{\text{VP}}{\text{VP + FN}} \tag{2.45}$$

- Especificidade - corresponde à taxa de verdadeiros negativos, ou seja, representa a capacidade do modelo predizer as observações que se encontram na categoria 0 da variável de interesse. De outra forma, representa a probabilidade de uma observação ser classificada como negativa dado que é, de facto, negativa. É dada por:

$$E = \frac{\text{Número de verdadeiros negativos}}{\text{Total de negativos}} = \frac{\text{VN}}{\text{VN + FP}} \tag{2.46}$$

- *Accuracy* - corresponde à proporção de predições corretas. Esta medida é altamente susceptível a conjuntos de dados desequilibrados, uma vez que não tem em consideração o número de elementos pertencentes a cada categoria. Como tal, facilmente pode conduzir a conclusões erradas sobre o desempenho do modelo. É dada por:

$$ACC = \frac{\text{Número de predições corretas}}{\text{Número total de observações}} = \frac{\text{VP + VN}}{\text{N}} \tag{2.47}$$

  onde $N$ corresponde ao número total de dados no conjunto.

- Eficiência - corresponde à média aritmética entre a sensibilidade e a especificidade. Na prática, a sensibilidade e a especificidade variam em direcções opostas na medida em que, geralmente, quando um modelo é muito sensível a positivos, tende a gerar muitos falsos positivos e vice-versa. Assim, um modelo de decisão perfeito (100% de sensibilidade e 100% de especificidade) raramente é alcançado. É dada por:

$$EFF = \frac{\text{S} + \text{E}}{2} \tag{2.48}$$

Uma outra medida que permite diagnosticar o modelo é o *Receiver Operating Characteristic Curve*, mais conhecida por ROC *Curve*. A curva ROC é um gráfico que ilustra o *tradeoff* entre a taxa de verdadeiros positivos e a taxa de falsos positivos, (designa-se como o complementar da especificidade, ou seja, $1 - E$) (Dangeti, 2017). A figura 2.1 ilustra três tipos de curva ROC, onde o eixo X representa a taxa de falsos positivos e o eixo Y a taxa de verdadeiros positivos. O ponto ideal na curva seria (0, 1), quer isto dizer, todas as observações positivas são classificadas correctamente e não existe uma observação negativa classificada como positiva (Rokach e Maimon, 2015).

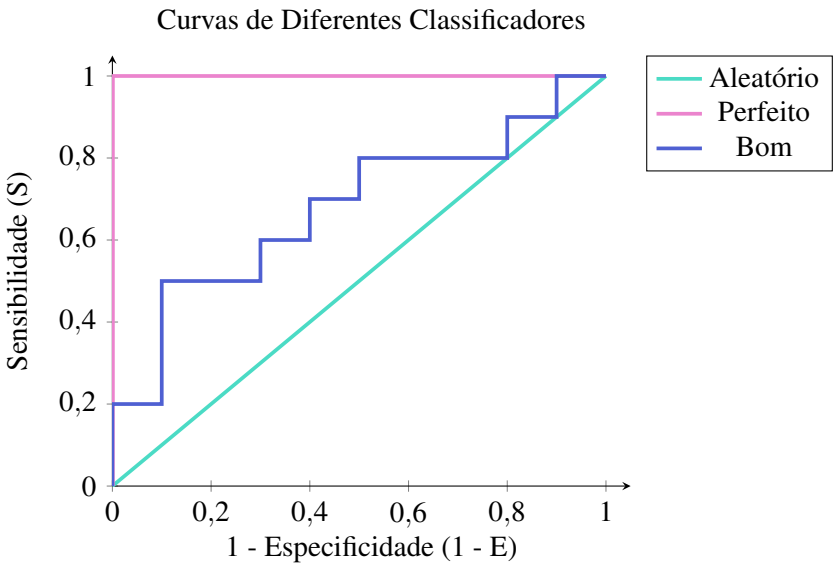



Figura 2.1: Curva ROC

Uma maneira simples de entender a utilidade da curva ROC é que, se o valor do *cut-off*, contido no intervalo [0, 1] for muito baixo, a maioria das observações previstas serão classificadas como positivas, mesmo quando algumas delas deveriam ser classificadas na categoria negativa. Por outro lado, estabelecer o valor do *cut-off* elevado penaliza a categoria de predição positiva. Idealmente, o valor do *cut-off* deve ser definido de maneira a compensar o valor entre as duas categorias e produzir maior precisão. Assim, o valor óptimo do *cut-off* corresponde ao valor do ponto de corte que permite o máximo da soma entre a sensibilidade e a especificidade (Dangeti, 2017).

A Área Abaixo da Curva/*Area Under the Curve* (AUC) é uma medida empírica de classificação da *performance* (Sammut e Webb, 2017). Esta medida varia entre no intervalo [0, 1]. Como tal, para cada valor de AUC é possível obter diversas conclusões sobre a *performance* do modelo, tabela 2.2.

Tabela 2.2: Interpretação dos valores de AUC (Andreozzi, 2012).

| **AUC** | **Diagnóstico** |
|---|---|
| 0,5 | Modelo sem poder discriminatório |
| $0{,}7 \leq$ AUC $< 0{,}8$ | Discriminação aceitável |
| $0{,}8 \leq$ AUC $< 0{,}9$ | Discriminação excelente |
| AUC $\geq 0{,}9$ | Discriminação extraordinária |

## 2.2 Árvores de Decisão

Uma das abordagens mais promissoras e populares no *data mining* é o uso das árvores de decisão. As árvores de decisão são técnicas simples, mas bem sucedidas, para prever e explicar a relação existente entre as variáveis preditoras e a variável de resposta. Além do uso do *data mining*, as árvores de decisão, que são derivadas da lógica, gestão e estatística, são hoje ferramentas altamente eficazes noutras áreas de negócio, como por exemplo no *text mining*, extracção de informação, aprendizagem automática e reconhecimento de padrões (Rokach e Maimon, 2015).

Uma árvore de decisão é um modelo preditivo que pode ser usado para representar os modelos de classificação, bem como os de regressão. O tomador de decisão, consoante o seu objectivo, identifica a estratégia que pretende adoptar. No âmbito deste projecto usar-se-á o método de classificação.

As árvores de classificação são usadas para classificar um objecto ou uma instância num conjunto predefinido de classes, com base nos valores dos seus atributos. Estas árvores são frequentemente usadas em áreas bastante interessantes para o negócio, nomeadamente finanças, *marketing*, engenharia e medicina. Esta metodologia é útil como uma técnica exploratória. No entanto, não substitui os métodos estatísticos tradicionais existentes e, como tal, existem muitas outras técnicas que podem ser usadas para classificar ou prever um conjunto de instâncias de um grupo predefinido de classes, como *neural networks* ou *support vector machines*.

Este algoritmo das árvores de decisão é um modelo de *machine learning* que pode ser aplicado a dados categóricos e contínuos. O conceito principal do algoritmo é dividir os dados consecutivamente de acordo com certos critérios. Assim, consiste num conjunto de nós que formam uma árvore enraizada, isto é, uma árvore com um nó chamado *root* que não possui arestas de entrada e, como tal, os restantes nós têm exactamente uma entrada. Um nó com arestas de saída é designado por um nó interno, enquanto, os restantes são chamados de folhas (também conhecidos como nós terminais).

Cada nó interno divide o espaço da instância em dois ou mais sub-espaços de acordo com uma determinada função discreta dos valores dos atributos de entrada. No caso mais simples e frequente, cada teste considera um único atributo, de modo que o espaço da instância seja particionado de acordo com o valor dos atributos. No caso de atributos numéricos, a condição refere-se a um intervalo (Rokach e Maimon, 2015).

As árvores de decisão são computacionalmente eficientes no treino e bastante rápidas para classificar as novas instâncias. Uma característica negativa das árvores de decisão é que são sensíveis ao excesso de ajustamento, *overfitting*, mas que pode ser contornado através do *pruning* ou da adição de outros critérios à construção da árvore (Chatzidimitriou et al., 2018). O processo de *pruning* consiste na simplificação da árvore de decisão gerada inicialmente.

Supondo o exemplo em que se pretende prever a compra de uma carteira feminina, a figura 2.2 representa uma árvore de decisão que incorpora atributos nominais e numéricos.

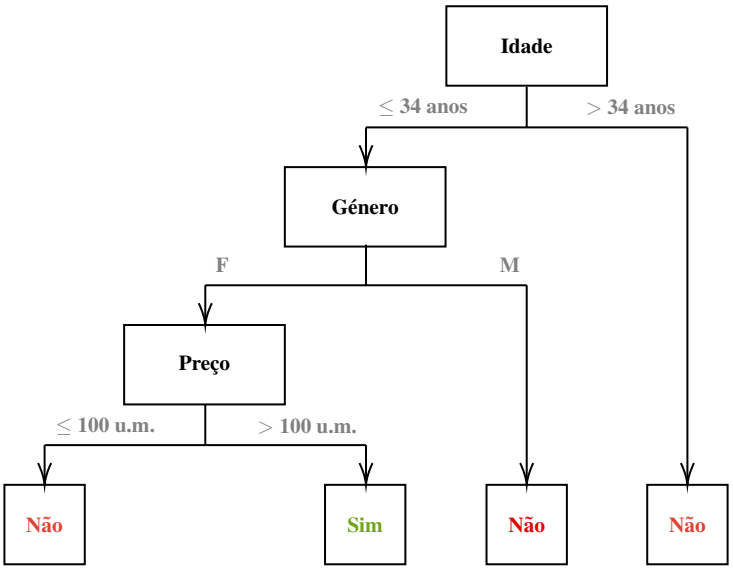



Figura 2.2: Exemplo de representação de uma árvore de decisão com três variáveis independentes.

Assim, o analista pode prever a resposta de um potencial comprador de uma carteira feminina e, por

sua vez, entender as características comportamentais dos potenciais clientes.

### 2.2.1 Medidas de Divisão

Na maioria das árvores de decisão, as funções de divisão discretas são univariadas, isto é, um nó interno é dividido de acordo com o valor de um único atributo. Consequentemente, o algoritmo procura o melhor atributo para executar a divisão. Existem vários critérios univariados que podem ser caracterizados de diferentes maneiras, como por exemplo (Rokach e Maimon, 2015):

- de acordo com a origem da medida de divisão: *Information Theory*, *Dependence* e *Distance*;
- de acordo com a estrutura da medida de divisão: *Impurity-based criteria*, *Normalized Impurity-based criteria* e *Binary criteria*.

No âmbito deste trabalho usar-se-ão medidas de *impurity*. As árvores de decisão dividem variáveis de forma recursiva, com base nos critérios de *impurity* definidos até atingirem alguns critérios de paragem, como por exemplo: observações mínimas por nó terminal e/ou observações mínimas para divisão em qualquer nó (Dangeti, 2017). Alguns dos critérios utilizados são:

- Entropia (*Entropy*): corresponde à medida da impureza nos dados. Se a amostra for completamente homogénea, a entropia será zero e, se a amostra for igualmente dividida, a entropia será um. Nas árvores de decisão, o preditor com maior heterogeneidade será considerado o mais próximo do nó raiz para classificar os dados fornecidos em classes num modo ganancioso. É dada por:

$$\text{Entropia} = -p_i \cdot \log_2 p_i, \tag{2.49}$$

  onde $i$, $i = 1,...,n$, corresponde ao número de classes e $p_i$ à probabilidade de escolher aleatoriamente um elemento da classe $i$, ou seja, a proporção do conjunto de dados da classe $i$. A entropia é máxima no meio, com o valor um, e mínima nos extremos com o valor zero. É desejável que o valor da entropia seja baixo, o que implica uma melhor agregação das classes.

- Ganho de Informação (*Information Gain*): corresponde à redução esperada na entropia causada pela nova divisão dos dados, de acordo com um determinado atributo. A ideia é começar com classes mistas e ir particionando de forma recursiva até que cada nó atinja as observações da classe mais pura. Em todas as etapas, a variável com ganho máximo de informação é escolhida de forma gananciosa. É dado por:

$$\text{Ganho de informação} = \text{Entropia do nó pai} - \sum_{i=1}^{k} w_i - \text{Entropia}_i, \tag{2.50}$$

  onde $k$ corresponde ao número de filhos do nó pai, e onde

$$w_i = \frac{n_i}{\sum_{i=1}^{k} n_i}, \tag{2.51}$$

  sendo que, $n_i$ diz respeito ao número de observações do filho $i$.

- *Gini*: corresponde à medida de classificação incorrecta. Esta medida funciona de maneira seme-

lhante à entropia, excepto que a medida *Gini* torna-se mais rápida de calcular:

$$Gini = 1 - \sum_i p_j^2, \tag{2.52}$$

onde $j$ corresponde ao número de classes.

A semelhança entre *gini* e a entropia é dada na figura 2.3.

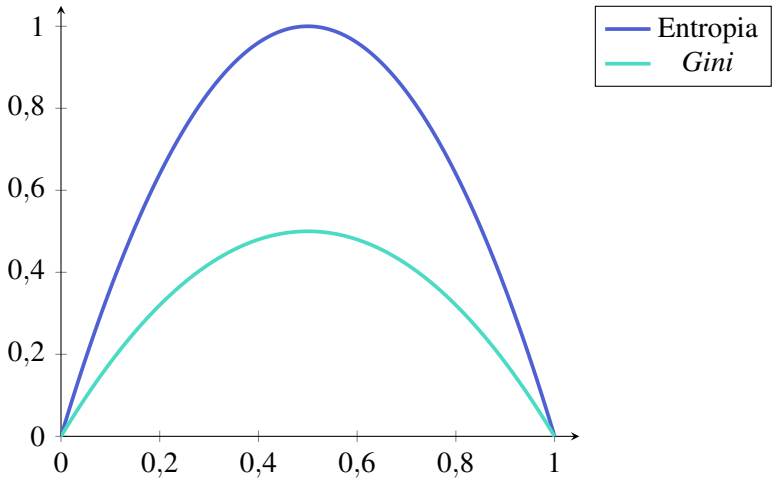



Figura 2.3: Relação existente entre as medidas *Gini* e Entropia.

É necessário ter em consideração que nos casos em que a variável *target* (resposta) é uma variável binária, a entropia é uma forma de medir a probabilidade de obter um elemento classificado como positivo a partir de uma selecção aleatória dos subconjuntos de dados. Assim, trata-se de uma medida de surpresa, se a entropia for zero não existem surpresas nas respostas possíveis.

Construir uma árvore de decisão passa por encontrar regras sobre as variáveis do modelo (ou pontos de corte) que retornam o maior ganho de informação, isto é, que tornam os ramos da árvore mais homogéneos e, por sua vez, com menor valor de entropia.

O uso de critérios de paragem bastante robustos tende a criar árvores de decisão pequenas e mal ajustadas. Por outro lado, o uso de critérios de paragem mais simples tende a gerar grandes árvores de decisão e, consequentemente, sobreajustadas ao conjunto de dados de treino. Para resolver este problema recorrente no uso das árvores de decisão, Breiman et al., 1984 desenvolveram uma metodologia de *pruning* baseada num critério de paragem que permite que a árvore de decisão seja sobreajustada ao conjunto de dados de treino. Posteriormente, a mesma árvore é "podada", tornando-a menor, removendo ramos que não contribuem para a sua precisão. Ou seja, a árvore pode continuar a crescer até que haja tantos nós quanto o número de conjuntos distintos de valores das variáveis explicativas. Para que uma árvore de classificação tenha um desempenho melhor para previsões futuras e, por sua vez, não ser *overfitted*, alguns ramos da árvore produzidos pelo algoritmo podem ser eliminados (Rokach e Maimon, 2015).

Uma motivação importante para o uso da metodologia de *pruning* é a troca da precisão pela simplicidade (Bratko e Bohanec, 1994). Quando o objectivo é produzir uma árvore de decisão devidamente precisa, o *pruning* traduz-se numa metodologia bastante útil. Uma árvore de decisão inicial é vista como uma árvore completamente precisa. Assim, uma árvore de decisão *pruned* indica o quão próxima esta está da árvore inicial.

A escolha de um parâmetro de complexidade $\lambda$ determina a extensão do *pruning*. Quando o valor do

parâmetro é igual a zero, obtém-se a árvore mais complexa possível. Um valor de $\lambda$ crescente, implica um maior *pruning* da árvore e, por sua vez, a árvore fica mais simples (Agresti, 2019).

### 2.2.2 Validação Cruzada

Depois dos dados estarem preparados é uma boa prática seleccionar uma estratégia de validação. A capacidade de generalização de um modelo a partir de uma dada amostra de dados pode ser fraca devido ao subajustamento (*underfitting*) ou ao sobreajustamento (*overfitting*).

O subasjustamento traduz-se no fraco ajustamento do modelo ao conjunto de dados e, é facilmente detectável a partir de medidas de qualidade de ajustamento (como o coeficiente de determinação - $R^2$). É esperado que um modelo com fraco ajustamento também tenha fracas previsões para novos conjuntos de dados.

O fenómeno de sobreajustamento surge quando o modelo tem um ajustamento aparentemente bom, mas na verdade não está a capturar uma boa generalização dos dados – está a ajustar-se ao ruído dos mesmos. Os erros de previsão são bastante superiores em valor absoluto aos erros de ajustamento quando este fenómeno está presente (EliteDataScience, 2019).

Para o âmbito deste trabalho usar-se-á o método de validação cruzada *k-fold*. Este método começa com a divisão da amostra em $k$ sub amostras (*folds*) de igual dimensão. Para amostras de grande dimensão, é usual o $k$ ser 10. No entanto, também é comum o $k$ ser igual a cinco. É necessário ter em consideração que a escolha das *folds* é aleatória. O algoritmo efectua $k$ iterações em que cada iteração $i$ tem o seguinte processo (Torrejano, 2018 e tabela (2.3)):

- é constituído um conjunto de treino que contém todas as *folds* excepto a *fold i*;
- o conjunto de teste ou de validação é constituído pela *fold i*;
- ajusta-se o modelo ao conjunto de treino;
- obtêm-se previsões de acordo com o modelo ajustado ao conjunto de treino para o conjunto de validação;
- a partir das previsões obtidas, são calculados os erros de previsão para as observações pertencentes ao conjunto de validação.

Tabela 2.3: Esquema do processo de uma validação cruzada de 3-*folds*.

| | **Sub amostra 1** | **Sub amostra 2** | **Sub amostra 3** |
|---|---|---|---|
| **Iteração 1** | **Teste** | Treino | Treino |
| **Iteração 2** | Treino | **Teste** | Treino |
| **Iteração 3** | Treino | Treino | **Teste** |

Ao caso particular do método de validação cruzada *k-fold* em que $k$=$n$ (número de observações na amostra) dá-se o nome de método *Leave-One-Out*.

### 2.2.3 Medidas de Erro

Uma vez treinado o modelo é necessário medir a capacidade discriminatória do mesmo. Este processo é idêntico ao que já foi referido na sub-secção 2.1.1.4.

Para além das medidas mencionadas anteriormente, existem outras medidas que são necessárias ter em consideração, nomeadamente (Rokach e Maimon, 2015):

- Precisão - corresponde à proporção de verdadeiros positivos entre o número de observações que são classificadas como positivas. É dada por:

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos}}{\text{Número de observações classificadas como positivas}} = \frac{\text{VP}}{\text{VP + FP}} \tag{2.53}$$

- F - corresponde à relação existente entre a Sensibilidade e a Precisão. Esta medida traduz-se numa medida mais realista de desempenho do modelo. Esta medida quando toma o valor zero, traduz-se em valores baixos de Sensibilidade e de Precisão. Analogamente, se os valores de Sensibilidade e de Precisão forem altos, então a medida F assume valor igual a um. É dada por:

$$\text{F} = 2 \times \frac{\text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} \tag{2.54}$$

  Pode-se observar na figura 2.4 a relação existente entre a Precisão e a Sensibilidade.

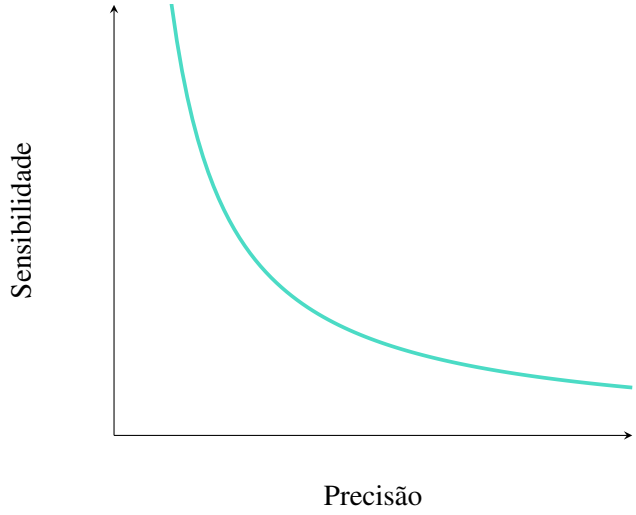



Figura 2.4: Relação existente entre as medidas Precisão e Sensibilidade

- Prevalência - corresponde à proporção de observações positivas. É dada por:

$$\text{Prevalência} = \frac{\text{VP} + \text{FN}}{N} \tag{2.55}$$

- Taxa de Falsos Positivos - corresponde à taxa de falsos positivos, ou seja, representa a capacidade do modelo predizer erradamente uma observação positiva. É dada por:

$$\text{TFP} = \frac{\text{FP}}{\text{FP} + \text{VN}} \tag{2.56}$$

- Taxa de Falsos Negativos - corresponde à taxa de falsos negativos, ou seja, representa a capacidade do modelo predizer erradamente uma observação negativa. É dada por:

$$\text{TFN} = \frac{\text{FN}}{\text{FN} + \text{VP}} \tag{2.57}$$

Desta forma, depois de explicadas as metodologias a utilizar neste projecto, no capítulo seguinte será feita uma descrição e uma análise exploratória das variáveis a utilizar da amostra, do período em análise. Por fim, serão apresentados os modelos obtidos para cada metodologia utilizada e, por sua vez, avaliar a capacidade discriminatória de cada modelo, de forma a obter o melhor modelo para este estudo.

# 3. Análise Exploratória de Dados

Neste capítulo será caracterizada a população em estudo e serão apresentados os métodos de análise exploratória utilizados no desenvolvimento do projecto. Ao longo da aplicação prática dos modelos apresentados na secção anterior ao caso de estudo foram encontradas algumas inconsistências ao nível do conjunto de dados utilizados. Sendo assim, serão também descritas algumas das dificuldades encontradas durante a implementação dos modelos.

## 3.1 Os Dados

Os dados utilizados para o presente trabalho são relativos a uma empresa portuguesa que integra o *Portuguese Stock Index* (PSI20). É composto pelas acções das 20 maiores empresas cotadas na bolsa de valores de Lisboa e reflecte a evolução dos preços das acções. Como tal, corresponde ao principal indicador de referência no mercado de capitais português. Actualmente, o PSI20 é composto por apenas 18 empresas (EURONEXT, 2018).

Ao abrigo do Regulamento Geral sobre a Protecção de Dados (RGPD), todos os dados e informações utilizados na realização deste projecto estão devidamente anonimizados, não sendo possível indicar o número de colaboradores da empresa. Este regulamento visa a protecção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados (A. d. República, 2019).

### 3.1.1 Limitações

Abaixo enumeram-se as principais limitações dos dados, que levaram a que fosse necessário excluir algumas observações:

1. Uma vez que existiam algumas incoerências nos dados, como por exemplo, a data de entrada na empresa ser superior à data de saída, eliminaram-se todas as observações que demonstravam inconsistência nos seus dados.

2. Para o *data set* em estudo, não foram tidos em consideração os casos jurídicos litigiosos existentes.

3. Por decisão da empresa, não se considera para este estudo os membros que constituem a equipa de gestão da entidade empresarial.

4. O Código do Trabalho compreende vários tipos de rescisão de contrato de trabalho e, como tal, para esta análise foram excluídos os colaboradores cujo motivo de saída é da responsabilidade da entidade patronal, isto é, saídas involuntárias. Esta decisão resume-se a um pressuposto da hipótese da seguinte abordagem: não se pode procurar um padrão na decisão de sair da empresa, se o colaborador não teve intervenção nessa decisão.

## 3.2 Abordagem ao Problema

Este estudo tem como objetivo contruir um modelo estatístico capaz de prever a saída de um colaborador de forma voluntária da empresa. Em primeiro lugar, é importante definir o que é considerado uma saída voluntária da empresa. Este conceito é de elevada importância para a realização deste projecto, dado que é a partir dele que o trabalho surge.

Segundo o Código de Trabalho, artigo 340º, (D. d. República, 2009) para além de outras modalidades legalmente previstas, o contrato de trabalho pode cessar por (Economias, 2016):

- Caducidade: considera-se que o contrato de trabalho caduca quando se verifica:
  1. o seu termo;
  2. a impossibilidade de o trabalhador prestar o seu trabalho ou de o empregador o receber;
  3. a reforma do trabalhador, por velhice ou invalidez.
- Revogação: esta modalidade acontece por acordo escrito entre o empregador e o funcionário.
- Despedimento por facto imputável ao trabalhador: trata-se de despedimento por iniciativa do empregador. Este acontecimento deriva de um comportamento culposo do colaborador, que pela sua gravidade e efeitos, torne imediatamente impossível a conservação da relação de trabalho.
- Despedimento colectivo: corresponde à cessação de contrato de trabalho promovida pelo empregador, abrangendo, pelo menos um grupo de 2 colaboradores, conforme a tipologia da dimensão da empresa. Refere-se também sempre que acontece o encerramento de uma ou várias secções ou a redução do número de trabalhadores.
- Despedimento por extinção de posto de trabalho: diz respeito à cessação de contrato de trabalho promovida pelo empregador fundamentada em motivos de mercado, estruturas ou tecnológicos, relativos à empresa.
- Despedimento por inadaptação: consiste no despedimento baseado na inadaptação superveniente do trabalhador ao posto de trabalho.
- Resolução pelo trabalhador: rescisão do contrato de trabalho por iniciativa do trabalhador, com ou sem justa causa.
- Denúncia pelo trabalhador: o trabalhador pode denunciar o contrato independentemente de justa causa, mediante comunicação ao empregador, por escrito, com a antecedência mínima de 30 ou 60 dias, conforme tenha, respectivamente, até 2 anos ou mais de 2 anos de antiguidade.

Assim, para este estudo e, consoante as políticas da empresa, a saída voluntária define-se como cessação do contrato de trabalho pela resolução e denúncia do trabalhador. De seguida, é necessário averiguar *a priori* quais serão as variáveis a incluir no modelo.

A empresa efectua entrevistas direccionadas aos colaboradores que se desvinculam da empresa, isto é, entrevistas de saída. Esta entrevista traduz-se numa conversa entre um colaborador da direcção de RH e o colaborador que manifesta interesse em rescindir contrato com a sua entidade patronal. Esta entrevista visa entender quais foram as principais razões que influenciaram o colaborador a tomar esta decisão. O modelo de entrevista de saída está disponível para consulta no Anexo A.

Através destas entrevistas, é possível saber quais são os padrões de resposta e, consequentemente, gerar *insights* sobre os principais motivos que influenciam a decisão dos colaboradores. Com isto, é

possível, *a priori*, escolher as variáveis a adicionar no modelo. Para determinar os padrões de resposta, aplica-se o teste de homogeneidade do Qui-Quadrado. Pretende-se testar se, para cada combinação existente entre os motivos que levaram à rescisão de contrato, existe homogeneidade nos padrões de resposta, consoante as métricas que identificam os indivíduos enquanto colaboradores.

Assim, com o objectivo de recolher extrair informação dos inquéritos realizados, foram efectuados os seguintes procedimentos:

1. Recolha de todos os resultados das entrevistas de saída.

2. Selecção das respostas à pergunta 11 - "*Em que medida os seguintes factores contribuíram para a tua decisão de sair da Empresa? Escala: 1 = contributo reduzido a 5 = enorme contributo.*".

3. Classificação das métricas utilizadas para testar se existem padrões de resposta, nomeadamente:

   - Idade: trata-se de uma variável quantitativa contínua, no entanto, para que seja possível incluí-la nesta abordagem, considera-se na forma de uma variável qualitativa ordinal, ou seja, $\leq 25$; 26 a 34; 35 a 44; 45 a 54 e, $\geq 55$ anos.
   - Antiguidade: trata-se de uma variável quantitativa contínua, porém, de forma a incluir esta variável nos padrões de resposta, é interpretada numa variável qualitativa ordinal, ou seja, $<$ 1; 1 a 2; 3 a 5; 6 a 10; 11 a 15; 16 a 20 e $\geq 21$ anos de antiguidade;
   - Grupo Organizacional: corresponde ao *job title* do colaborador, consoante a pirâmide da organização;
   - Área Funcional: refere-se à área em que o colaborador se insere no meio empresarial;
   - Avaliação de Desempenho: corresponde ao grupo de avaliação da *performance* de cada colaborador, nomeadamente: *High Performers*, *On Target* e *Low Performers*.

4. Agrupamento dos cinco níveis diferentes de resposta em apenas três categorias distintas: baixo, constituído pelos níveis 1 e 2; médio, pelo nível 3 e elevado pelos níveis de resposta 4 e 5.

5. Realização do teste de homogeneidade do Qui-Quadrado para o número de combinações possíveis entre os motivos de saída e as métricas utilizadas. O *script* utilizado para este algoritmo pode ser consultado no apêndice A.

6. Interpretação dos valores obtidos, tanto do *p-value* como do número de observações em cada padrão de resposta. Para a obtenção dos padrões de resposta, serão tidos em consideração os grupos compostos pelo menos por 25 colaboradores com um *p-value* superior a 0,05. Para esta abordagem, interpretam-se apenas os resultados cuja proporção de respostas de importância elevada é superior a 45%.

A título de exemplo, considere-se o seguinte caso: a empresa Lykke constituída por 1000 colaboradores, utiliza o mesmo modelo de carreira e de desenvolvimento que a empresa em estudo e efectua o mesmo procedimento com as entrevistas de saída. Assim sendo, a empresa Lykke identifica os colaboradores a partir das métricas mencionadas acima. Sabe-se que saíram da empresa, num espaço de dois anos, cerca de 200 colaboradores. Nesta amostra, apenas 125 correspondem a saídas voluntárias.

Considere-se o universo amostral constituído pelos colaboradores que rescindiram contrato. De seguida, divide-se em sub-amostras consoante as métricas. Por exemplo, inicia-se a primeira divisão através da métrica Idade e de seguida, uma segunda divisão para os indivíduos cuja idade é inferior a 25 anos.

Posto isto, para cada combinação entre os motivos de saída efectua-se o teste de homogeneidade do Qui-Quadrado. Considere-se que a primeira experiência corresponde ao conjunto das respostas aos motivos ”Oportunidades de Desempenho” e ”Formação”. Pretende-se averiguar se existe um padrão de homogeneidade nas respostas aos motivos seleccionados, para o sub-grupo da métrica escolhida inicialmente. De seguida, efectua-se a mesma experiência para os motivos ”Oportunidades de Desenvolvimento” e ”Compensação”, e assim sucessivamente.

Regressando ao caso de estudo deste projecto, os resultados obtidos desta análise podem ser observados na figura 3.1.

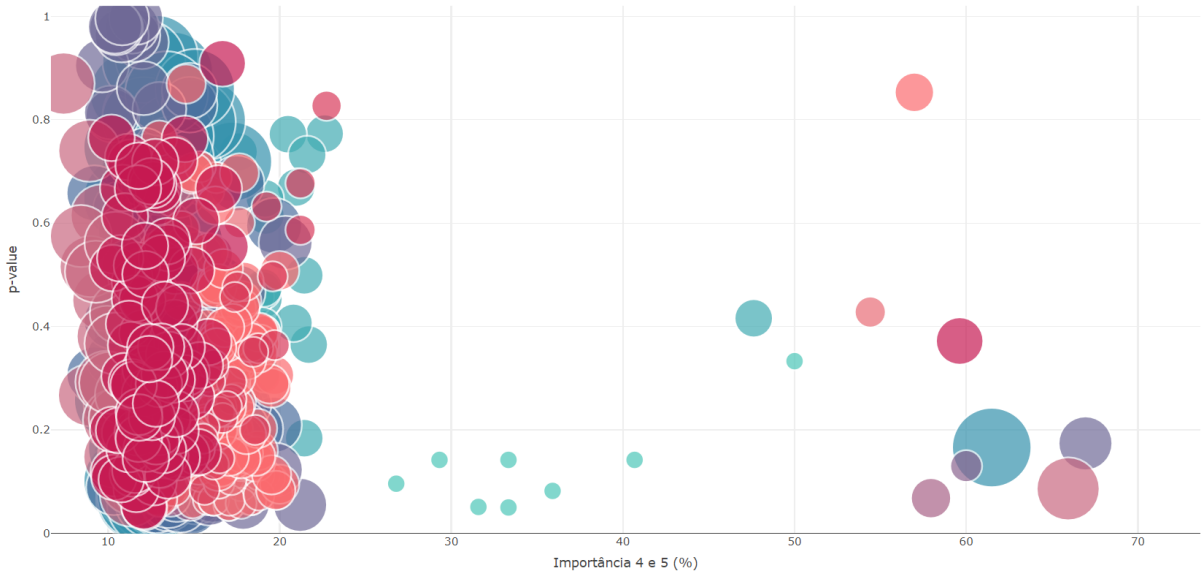



Figura 3.1: Comparação entre o *p-value* e o grau de importância.

De forma a facilitar a interpretação da figura 3.1, a dimensão de cada círculo da figura corresponde ao número de colaboradores que se inserem no respectivo padrão de resposta. O eixo das abcissas corresponde à percentagem de resposta do nível elevado para o grupo do padrão de resposta e, por sua vez, o eixo das ordenadas corresponde ao *p-value* associado ao teste de homogeneidade do Qui-Quadrado.

No entanto, de forma a não induzir em erro as conclusões obtidas através da visualização da figura 3.1, restringe-se a zona de observação do gráfico. Considerando o nível de significância mais usual, tendo em conta o estado da arte de HRA, de $\alpha = 0{,}05$, não existe evidência para afirmar que os padrões assinalados na figura 3.2 não seguem um padrão de homogeneidade. Restringe-se ainda mais a figura para os casos em que a proporção das respostas elevadas é superior a 45%.


Figura 3.2: Comparação entre o *p-value* ($>$ 0,05%) e o grau de importância.

Como se pode verificar a partir da figura 3.2, os principais motivos que levam o colaborador a rescindir o contrato de trabalho são, nomeadamente, as oportunidades de desenvolvimento e a compensação. Posto isto, estes dois motivos serão tidos em consideração nas variáveis a incluir no modelo.

O resultado desta predição estatística sobre a saída voluntária do colaborador tem o propósito de aumentar a retenção dos funcionários com a melhor *performance*, por parte da empresa do estudo.

## 3.3 As Variáveis em Estudo

Neste projecto as variáveis estão caracterizadas em quatro grupos distintos, consoante o meio ambiente em que se inserem, nomeadamente: sociodemográfico, desempenho e desenvolvimento, socioeconómico e exógeno. Apresentam-se ainda algumas variáveis que foram excluídas do modelo.

### 3.3.1 Variáveis sociodemográficas

As variáveis sociodemográficas referem-se às que correspondem ao indivíduo e, por sua vez, dependem unicamente do mesmo.

#### Status

Corresponde à variável fulcral para a construção de qualquer modelo no desenvolvimento deste trabalho. Esta variável traduz-se no *status* do colaborador no período em análise. Trata-se de uma variável qualitativa nominal, em que:

$$Status = \begin{cases} 1, & \text{se o colaborador saiu da empresa de forma voluntária} \\ 0, & \text{se o colaborador permaneceu na empresa} \end{cases}$$

### Género

Devido às desigualdades de género existentes no mercado de trabalho, é de certa forma paradoxal que seja o género feminino aquele que detém uma maior satisfação no trabalho face ao género masculino, porém corresponde ao género mais propenso à rotatividade laboral voluntária. Torna-se interessante o facto de que, o género feminino representa uma maior taxa de rotatividade, especialmente no início do seu percurso profissional. Contudo, quando as características pessoais e profissionais são analisadas em conjunto, a diferença entre as propensões em relação à saída voluntária entre cada género são quase nulas. É necessário ter em consideração que as taxas mais elevadas de *turnover* no género feminino são fortemente relacionadas com a família e a gravidez (Lee, 2012).

### Estado civil

Segundo Ahituv e Lerman, 2005, a estabilidade laboral influencia o estado civil do indivíduo e vice-versa. O estudo reforça a ideia de que existe uma forte evidência de que a instabilidade no percurso profissional reduz a probabilidade de matrimónio. Ao mesmo tempo, o matrimónio aumenta a estabilidade no emprego.

### Idade

A geração *millennials*, também conhecida por geração Y, refere-se a um conceito em Sociologia que corresponde ao grupo dos indivíduos que nasceram entre os inícios da década de 1980 e da década de 2000. Os *millennials* têm sido estigmatizados como *job-hoppers*, ou seja, é uma geração que é caracterizada através da elevada frequência de mudança de emprego. Segundo Gregory, 2019, 75% desta geração acredita que a mudança de emprego é um ponto a favor na sua carreira profissional. É importante destacar que, no entanto, o próprio estereótipo da geração é injustificado e a natureza da rotatividade laboral é mais relacionada com a idade do que com a dinâmica da sua geração.

### Número de dependentes

Actualmente, um dos factores mais importantes para um colaborador se sentir motivado num espaço laboral é conseguir manter o equilíbrio entre a vida profissional e a familiar. Para efeitos de integração no agregado familiar, entende-se como dependente (Notícias, 2011):

- filhos, adoptados e enteados, menores não emancipados e menores sob tutela;
- filhos, adoptados, enteados e ex-tutelados, maiores, que, não tendo mais de 25 anos não auferem rendimentos anuais 14 vezes o salário mínimo nacional;
- filhos, adoptados, enteados e ex-tutelados, maiores considerados inaptos para o trabalho.

### Número de filhos

Esta variável pode induzir em erro, dado que traduz a mesma informação que a variável referente ao número de dependentes. No entanto, é necessário ter em consideração que um filho pode ou não ser considerado dependente e vice-versa.

#### Antiguidade

Um nível de antiguidade baixo poderá estar associado a uma maior probabilidade de saída voluntária (Dostie, 2005). No entanto, é necessário ter em consideração que um colaborador para poder avançar com o processo de rescisão de contrato, é necessário comunicar à entidade patronal com antecedência mínima de 30 ou 60 dias, conforme tenha, respectivamente, até dois anos ou mais de dois anos de antiguidade na empresa. Ou seja, quanto maior for a antiguidade na empresa, maiores serão os encargos necessários para a rescisão do contrato, tanto para a empresa como para o colaborador.

#### Local de trabalho

A empresa é sediada na zona A e é composta por diversos edifícios em Portugal Continental, Ilhas e outros territórios. Assim, consoante a localização do edifício, pode-se estar presente em vários ambientes de trabalho. Como tal, esta variável divide-se em três grandes grupos: A, B e C.

#### Tipo de contrato

Existem dois tipos de contratos de trabalho, nomeadamente, o contrato de trabalho a termo certo e contrato de trabalho a termo incerto. Esta variável torna-se interessante neste estudo, na medida em que permite avaliar se a segurança de um contrato sem termo tem impacto na decisão de saída do colaborador de forma voluntária.

#### GO

O GO, denominado por grupo organizacional, representa o cargo que o colaborador desempenha na empresa, isto é, o *job title* associado às responsabilidades e à experiência de cada colaborador.

### 3.3.2 Variáveis de desempenho e desenvolvimento

As variáveis de desempenho e desenvolvimento correspondem à relação existente entre a empresa e o colaborador, uma vez que, tanto depende da *performance* do colaborador como da conjugação de diversos factores inerentes à empresa, nomeadamente: do modelo de carreiras, dos objectivos e dos valores da mesma.

#### Avaliação de desempenho

Como o objectivo é prever a saída de um colaborador de forma voluntária, consequentemente, a empresa tenciona reter os colaboradores que detêm um melhor nível de desempenho. Actualmente, o custo associado à contratação de talentos é superior ao custo associado à retenção do colaborador com elevado nível de desempenho (Lewis, 2018).

#### Número de mobilidades

Entende-se por mobilidade a mudança de área de negócio dentro da própria empresa. O processo de mobilidade pode ser iniciado pela manifestação de interesse do colaborador em novas experiências ou pela adequação do perfil do mesmo a um novo desafio organizacional. No entanto, nem todas as mobilidades estão directamente associadas a um pedido de um colaborador.

#### Número de promoções funcionais

As promoções funcionais caracterizam-se pela alteração do *job title* do colaborador, isto é, um colaborador é classificado como júnior na empresa e, derivado da promoção funcional, o *job title* do colaborador altera-se para, por exemplo, sénior. É importante referir que nem sempre as promoções funcionais estão relacionadas com promoções salariais. As promoções funcionais seguem o modelo de carreiras estabelecido pela empresa em estudo.

#### Horizonte temporal das promoções funcionais

Entende-se como horizonte temporal, em função das promoções funcionais, a diferença temporal entre o momento da extracção da base de dados e o último momento em que se proporcionou uma promoção funcional.

A título de exemplo, o colaborador ABC entrou na empresa no dia 1 de Dezembro de 2017 e a extracção dos dados foi realizada no dia 31 de Janeiro de 2018. Durante este período o colaborador não esteve sujeito a nenhuma promoção funcional, portanto o valor da variável é a diferença temporal entre o dia da extracção dos dados e o dia de entrada na empresa, ou seja, a sua antiguidade.

#### Horizonte temporal das mobilidades

Entende-se como horizonte temporal, em função das mobilidades, a diferença temporal entre o período definido para o início da análise e o último momento em que o colaborador esteve sujeito a uma mobilidade.

Por exemplo, o colaborador YWZ entrou na empresa no dia 1 de Março de 2016 e o período definido para o início da análise foi o dia 31 de Janeiro de 2018. Durante este período o colaborador esteve sujeito a duas mobilidades, a 1 de Maio de 2017 e a 2 de Janeiro de 2018. Assim, o valor da variável corresponde à diferença de dias entre o dia da extracção dos dados e o dia em que se celebrou o último momento da mobilidade, ou seja, 0,08 anos.

### 3.3.3 Variáveis socioeconómicas

Relativamente às variáveis socioeconómicas, representam o custo de aquisição e manutenção de recursos humanos em contrapartida aos serviços prestados pelo colaborador.

#### Retribuição anual salarial

Segundo a *Society for Human Resource Management*, em 2008 foi efectuado um estudo sobre a satisfação no trabalho, em que 92% dos colaboradores afirmaram que o valor da remuneração influenciava a sua satisfação. Os gestores das pequenas empresas não devem subestimar o efeito que o salário tem na retenção de empregos (Duggan, 2020).

#### Número de promoções salariais

As promoções salariais são caracterizadas pela alteração do vencimento mensal do colaborador. Esta variável está relacionada com os resultados obtidos consoante os objectivos individuais estabelecidos para cada colaborador.

#### Horizonte temporal das promoções salariais

Entende-se como horizonte temporal, em função das promoções salariais, a diferença temporal entre o momento da extracção da base de dados e o último momento em que se proporcionou uma promoção salarial. Caso o colaborador não tenha nenhuma promoção salarial, esta variável toma o valor associado à antiguidade. A forma de interpretar esta variável é idêntica às variáveis apresentadas no grupo 3.3.2.

#### % de aumento salarial

Esta variável refere-se à percentagem de aumento salarial face ao último momento em que se proporcionou uma promoção salarial. Caso o colaborador não tenha nenhuma promoção salarial, esta variável toma o valor zero.

#### Diferença salarial face ao *target* do GO

Como já foi referido anteriormente, a cada colaborador é associado um GO e, por sua vez, a empresa estabelece uma banda salarial consoante o *job title*. Sendo assim, é calculado a distância ao *target* salarial do seu grupo organizacional.

### 3.3.4 Variáveis exógenas

O grupo das variáveis exógenas, como o nome indica, refere-se ao conjunto das variáveis externas ao meio ambiente onde se estabelece a relação colaborador e empresa e, como tal, não dependem das decisões da mesma nem do comportamento do colaborador.

#### Diferença salarial face ao mercado de referência

Uma vez que a empresa pertence a um determinado sector económico, torna-se interessante para o projecto avaliar a diferença salarial de cada colaborador, consoante o seu *job title*, relativamente à mediana do sector.

#### Diferença salarial face ao mercado nacional

Dado que o estudo se refere a uma empresa nacional, é apelativo para a análise a variável que se refere à diferença salarial face ao mercado nacional, por cada *job title*.

### 3.3.5 Limitações na escolha de variáveis

Por fim, nem sempre se pode incluir todas as variáveis no modelo por diversas razões, como por exemplo, a falta de recursos necessários para recolha de informações.

#### Dimensão da equipa

A dimensão da equipa poderá ser um motivo relevante para a saída voluntária do colaborador. Uma vez que se trata de uma empresa que se encontra num processo de reestruturação organizacional, não seria viável recolher informação relativa à dimensão da equipa. Desta forma, não se consegue, de forma cuidada e precisa, obter a dimensão de cada equipa.

### Chefia directa

Como já foi referido anteriormente, nas entrevistas de saída um dos possíveis factores que poderão levar à saída voluntária dos colaboradores refere-se à relação existente com as chefias. No entanto, esta variável não poderá ser usada no modelo, dado que é composta por mais de 50 opções de chefia directa o que, por sua vez, dificulta a interpretação na metodologia a implementar.

### Satisfação no trabalho

No contexto da psicologia do trabalho, a satisfação no trabalho é a atitude geral do colaborador face ao seu trabalho e depende de vários factores psicossociais. A recolha dos dados para esta variável requer recursos que a empresa, actualmente, não possui. Para além da dificuldade de obtenção destes dados, existe sempre um risco associado à fiabilidade dos valores obtidos, devido à possibilidade dos inquiridos não responderem com a maior sinceridade.

### Pesquisa sobre o processo de saída

Um colaborador, antes de tomar qualquer decisão sobre o seu processo de saída, informa-se sobre os procedimentos necessários para efectuar a saída. Admitindo que um colaborador pesquisa sobre o processo de saída durante o seu horário de trabalho, ao usar o computador da empresa, esta detém qualquer informação acedida durante a sua utilização. Assim, seria possível reter os dados sobre esta variável, no entanto, a entidade patronal não detém os recursos suficientes para a obtenção da mesma.

### *Download* do recibo de vencimento

Como já foi referido anteriormente, se o colaborador utiliza o computador da propriedade da empresa, esta tem a capacidade de reter quais foram as pesquisas e os *download* efectuados durante a sua utilização. Se um colaborador tem por hábito fazer o *download* do recibo de vencimento de forma mensal, existe um padrão associado a esta pessoa. Caso contrário, se for um colaborador que não tem por hábito efectuar o *download* do recibo de vencimento e, ocasionalmente, guardou o ficheiro respectivo, como consequência, alterou-se o padrão do comportamento. A recolha dos dados desta variável traduz-se num processo exaustivo para a entidade patronal e, mais uma vez, esta não detém os recursos suficientes para sua recolha.

### Número de posições anteriores do colaborador

Supõe-se que se um colaborador tem poucos anos de experiência profissional, mas possui já um número elevado de empregos, este será mais propenso a que, num futuro próximo, rescinda o seu contrato de trabalho. Porém, ainda existe algum preconceito com os profissionais que alteram o seu emprego de forma constante ou, por exemplo, os indivíduos que num ano já tiveram pelo menos cinco empregos distintos.

Dependendo da política de recrutamento da empresa, esta alteração de emprego constante poderá demonstrar alguma insegurança para a entidade patronal. Todavia, é da responsabilidade da empresa perceber quais foram as razões que levaram o colaborador a efectuar as diversas mudanças. Para incluir esta variável no modelo, seria necessário averiguar um colaborador de cada vez e, por sua vez, estudar o seu histórico profissional. Este processo seria bastante moroso e exaustivo devido ao detalhe que lhe é exigido.

### Oportunidades de trabalho

As grandes ofertas de emprego em Portugal estão localizadas sobretudo nas grandes e médias cidades do litoral do país e abrangem as mais diversas áreas. Uma vez que cerca de 60% da amostra trabalha na região de Lisboa e vários sectores de actividade se destacam na capital portuguesa, a rotatividade de emprego poderá ser bastante superior em Lisboa face a cidades mais pequenas. Esta variável não é incluída no modelo, visto que o principal objectivo deste estudo é avaliar as componentes internas que proporcionam a saída do colaborador de forma voluntária.

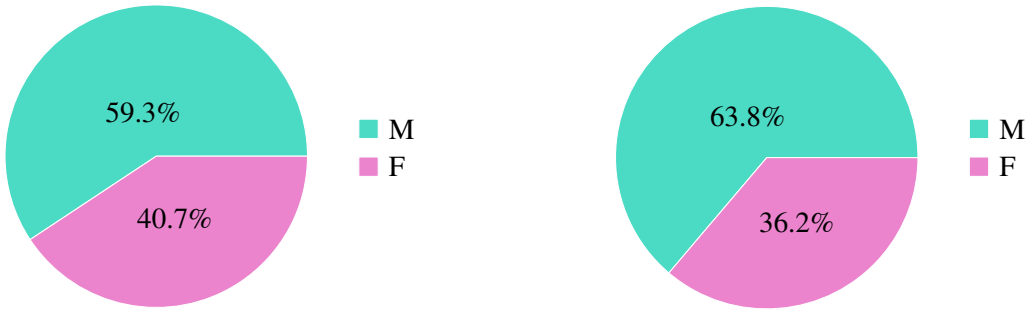
### Distância entre o trabalho e casa

Segundo Blake, 2020, quando se vive muito longe do local de trabalho, o percurso que é necessário efectuar acaba por influenciar todos os aspectos da vida, tanto profissional como pessoal. Certamente, trabalhar longe de casa acarretará custos significativamente mais elevados, tanto a nível financeiro como emocional. Estas consequências derivadas da distância entre o local de trabalho e a casa, reflectem-se nas decisões em relação ao processo de saída.

## 3.4 Análise de Dados

Para qualquer análise efectuada neste trabalho, foi garantido o número mínimo de 100 colaboradores, de forma a assegurar resultados sólidos e confiáveis.

A primeira representação gráfica dos dados dos colaboradores obtida foram gráficos circulares relativos à divisão entre os géneros, tanto para a amostra em análise, como para a amostra das saídas voluntárias. Como se pode verificar a partir da figura 3.3a, cerca de 59,3% da organização é constituída pelo género masculino, o que reflecte a área de negócio da empresa, onde existe uma predominância de áreas técnicas, tipicamente mais procuradas pelo género masculino (Empresa, 2018).




(a) Activos.

(b) Saídas Voluntárias.

Figura 3.3: Distribuição dos colaboradores pelo género.

Comparativamente à distribuição das saídas voluntárias, pode-se verificar que, cerca de 64% corresponde às saídas do género masculino. Como já foi referido anteriormente, a área de negócio da empresa é predominada pelo género masculino. Portanto, espera-se que dada a competitividade do mercado de trabalho na área de negócio, o volume de saídas voluntárias do género masculino seja superior à do género feminino.

A distribuição da faixa etária reflecte a aposta crescente na integração e desenvolvimento de jovens, o que se traduz nas políticas de recrutamento da empresa. Estas políticas baseiam-se no recrutamento

de jovens recém-licenciados para integrar diferentes áreas de negócios, assim como na oferta de estágios profissionais, estágios de verão e estágios curriculares (Empresa, 2018).

Tabela 3.1: Características amostrais da variável idade, consoante a amostra em estudo.

| Medida | Mínimo | Máximo | $Q_{\frac{1}{4}}$ | $Q_{\frac{1}{2}}$ | $Q_{\frac{3}{4}}$ | Média | Desvio Padrão |
|---|---|---|---|---|---|---|---|
| Activos | 21 | 63 | 37 | 42 | 46 | 41,12 | 7,41 |
| Saídas Voluntárias | 21 | 66 | 27 | 32 | 38 | 32,76 | 6,76 |

A tabela 3.1 apresenta as principais características amostrais da distribuição da idade da amostra em estudo. A grande maioria, 75% dos colaboradores activos, é composta por indivíduos cuja idade é inferior a 46 anos. Comparativamente aos colaboradores que rescindiram contrato com a empresa de forma voluntária, cerca de 75% têm até 38 anos. Tanto na amostra relativa aos membros activos como aos membros que constituem a amostra das saídas voluntárias, o valor da média e da mediana são iguais, podendo concluir-se que se trata de amostras simétricas.

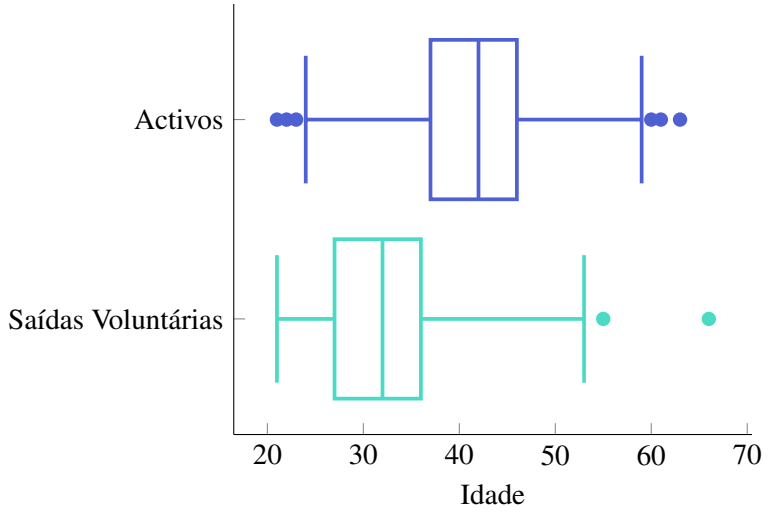



Figura 3.4: *Boxplots* paralelos da variável idade dos colaboradores (Activos e Saídas Voluntárias).

A representação em *boxplot* de uma amostra, reflecte como é a distribuição de uma variável em estudo. Este diagrama é muito útil para identificar assimetrias nos dados, caso a "caixa" esteja partida em dois pedaços muito diferentes, e para identificar possíveis candidatos a *outlier* na população. Entende-se por *outlier* uma observação que se destaca por ser muito extrema, ou seja, muito distante das restantes observações.

As conclusões que se podem retirar através da figura 3.4 correspondem às principais características amostrais, já calculadas na tabela 3.1. Para além da informação mencionada na tabela, pode-se verificar que existem candidatos a *outlier*. Relativamente à amostra correspondente aos colaboradores activos, a partir da representação gráfica, é possível identificar seis candidatos a *outlier* (21, 22, 23, 60, 61 e 63 anos). Estas observações correspondem a 21 indivíduos da amostra.

Comparativamente à amostra correspondente às saídas voluntárias, pode-se verificar apenas dois candidatos a *outlier* (55 e 66 anos). Esta observações correspondem unicamente a dois indivíduos da amostra.

A antiguidade dos colaboradores reflecte o vínculo do colaborador à empresa. É interessante para este estudo o facto de que, em média, os colaboradores demoram 6,11 anos até rescindirem contrato

de forma voluntária (tabela 3.2). No entanto, pode-se afirmar que existem colaboradores que anulam o contrato de forma voluntária em menos de um mês de contrato.

Tabela 3.2: Características amostrais da variável antiguidade, consoante a amostra em estudo.

| Medida | Mínimo | Máximo | $\mathcal{Q}_{\frac{1}{4}}$ | $\mathcal{Q}_{\frac{1}{2}}$ | $\mathcal{Q}_{\frac{3}{4}}$ | Média | Desvio Padrão |
|---|---|---|---|---|---|---|---|
| Activos | 0 | 35,6 | 7,6 | 13,6 | 18,62 | 12,9 | 7,11 |
| Saídas Voluntárias | 0 | 23,3 | 1,95 | 4,8 | 8,5 | 6,11 | 4,98 |

A partir da figura 3.5 pode-se verificar que o valor do $\mathcal{Q}_{\frac{3}{4}}$ referente à amostra das saídas voluntárias é bastante próximo do $\mathcal{Q}_{\frac{1}{4}}$ da amostra dos colaboradores activos, cuja diferença é de 0,9 anos. Em ambas as amostras, é possível concluir que o valor do mínimo é igual a zero.

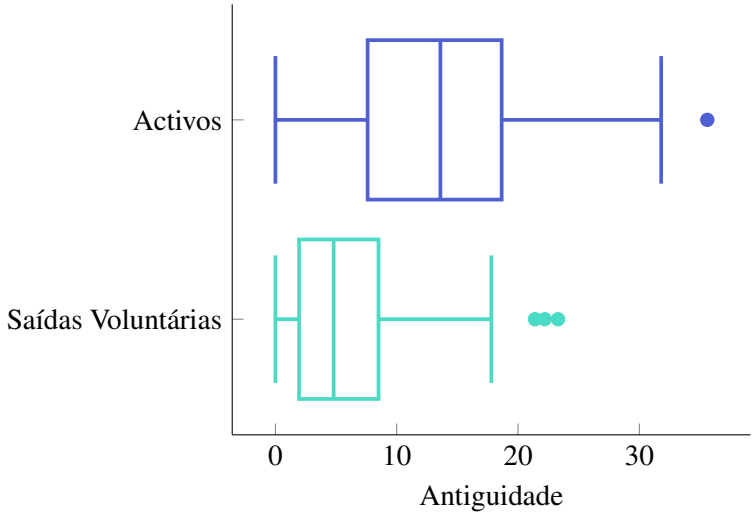



Figura 3.5: *Boxplots* paralelos da variável antiguidade dos colaboradores (Activos e Saídas Voluntárias).

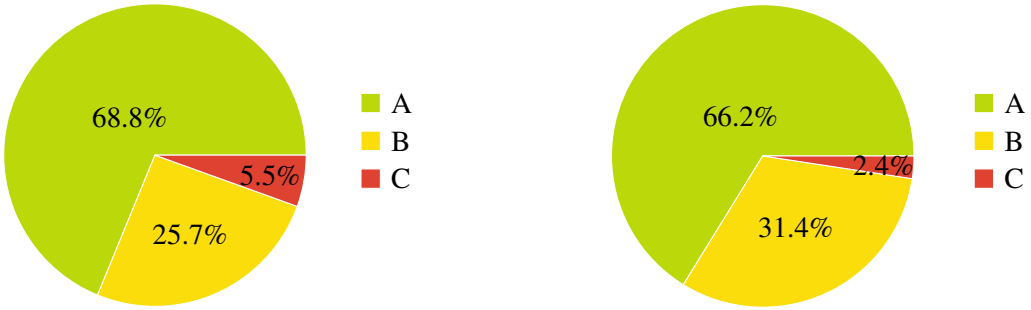
Relativamente à dispersão das observações, pode-se concluir, através da distância entre as barreiras inferior e superior, que a amostra dos colaboradores activos apresenta uma maior dispersão.

Ao contrário do que se podia verificar na variável idade, não se pode assumir que as amostras relativas à variável antiguidade são simétricas. Em relação à amostra relativa aos activos, como o valor da média é inferior ao valor da mediana, isto é, $\tilde{x} < \mathcal{Q}_{\frac{1}{2}}$, pode-se concluir que se está perante uma assimetria à esquerda. Em relação à amostra dos indivíduos que rescindiram contrato de forma voluntária, como o valor da média é superior ao valor da mediana, ou seja, $\tilde{x} > \mathcal{Q}_{\frac{1}{2}}$, conclui-se que se trata de uma assimetria à direita.

Relativamente à amostra correspondente aos colaboradores activos, é possível identificar um candidato a *outlier*. Esta observação diz respeito a apenas um indivíduo com antiguidade igual a 35,6 anos. Comparativamente à amostra correspondente às saídas voluntárias, verificam-se apenas três candidatos a *outlier* (21,4, 22,2, 23,3), que correspondem a três indivíduos da amostra.

A maior parte das operações da empresa em estudo encontra-se na área A, concentrando cerca de 67% dos colaboradores. É importante referir que a empresa detém também operações de igual importância na área B. Os restantes colaboradores, cerca de 5,5%, encontram-se dispersos por outras localizações (Empresa, 2018), como se pode observar através da figura 3.6a.

Relativamente às saídas voluntárias, como já foi referido anteriormente, as grandes ofertas de emprego em Portugal situam-se nas grandes cidades. Para além deste factor, o tecido empresarial em Portu-

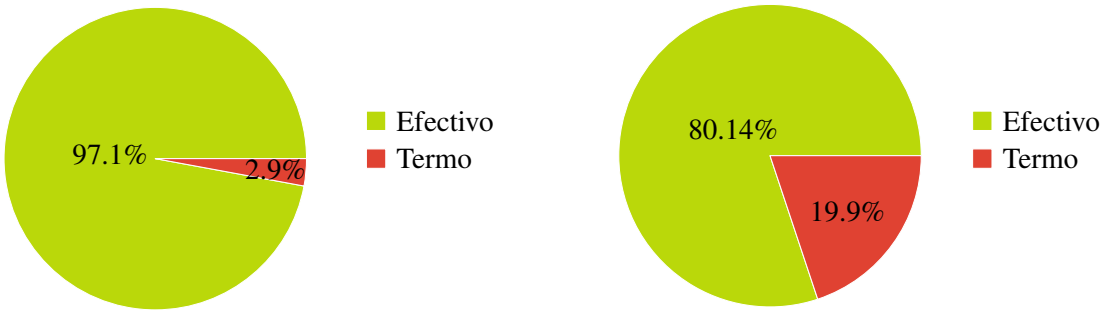



(a) Activos. (b) Saídas Voluntárias.

Figura 3.6: Distribuição dos colaboradores pelo local de trabalho.

gal está concentrado maioritariamente em Lisboa e no Porto e, por sua vez, espera-se que as saídas dos colaboradores tenham maior impacto nestas duas regiões metropolitanas. Através da figura 3.6b, pode-se verificar a distribuição das saídas voluntárias consoante as regiões A, B e C.

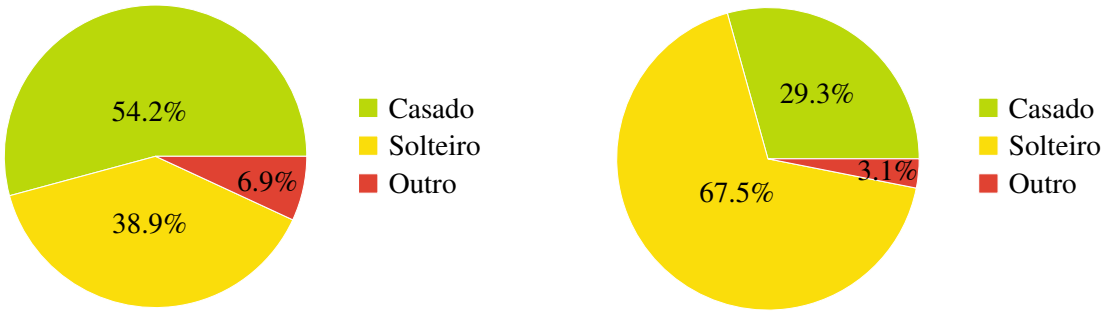
O compromisso com as políticas de empregabilidade sustentáveis é traduzido pela efectividade dos colaboradores: cerca de 97% dos membros da organização possui contrato efectivo, como se pode verificar através da figura 3.7. Para além disso, todos os colaboradores da empresa desempenham as suas funções a tempo inteiro (Empresa, 2018).




(a) Activos. (b) Saídas Voluntárias.

Figura 3.7: Distribuição dos colaboradores pelo contrato de trabalho.

Os gráficos apresentados na figura 3.8, descrevem a constituição da amostra consoante o estado civil dos colaboradores, referente aos activos e saídas voluntárias, figura 3.8a e figura 3.8b, respectivamente.




(a) Activos. (b) Saídas Voluntárias.

Figura 3.8: Distribuição dos colaboradores pelo estado civil.

A grande maioria dos colaboradores activos são casados, cerca de 54,2% da amostra, enquanto que 38.9% da mesma não possui qualquer tipo de compromisso matrimonial. No entanto, torna-se interes-

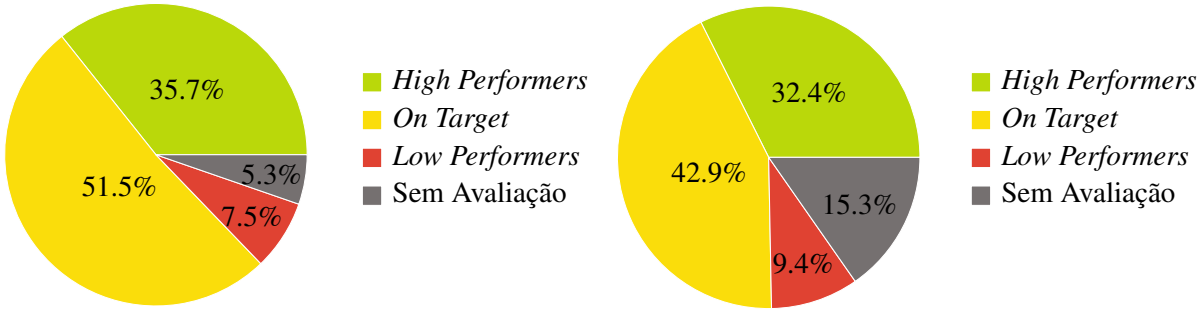
sante o facto de que, relativamente à amostra das saídas voluntárias, 3.8b, cerca de 68% dos indivíduos são solteiros. Como já foi mencionado anteriormente, segundo Ahituv e Lerman, 2005, existe uma relação entre o percurso profissional e o estado civil. Esta relação traduz-se numa maior probabilidade de vínculo do colaborador à empresa, caso este seja casado.

Relativamente à variável avaliação de desempenho, é importante categorizar a nota individual do colaborador, consoante as políticas da empresa em estudo. A partir da tabela 3.3, pode-se verificar que a nota quantitativa do colaborador pode variar entre 0% e 120%, consoante a realização dos objectivos predefinidos. Por decisão da empresa e para este estudo, foi tido em consideração apenas três grupos de classificação individual: *Low Performers* corresponde ao grupo de indivíduos cuja nota individual é inferior a 94%, *On Target* refere-se aos indivíduos cuja nota qualitativa pertence ao intervalo entre 95% e 105% e, por fim, *High Performers* diz respeito ao grupo cuja nota individual é superior ou igual a 106%.

Tabela 3.3: Classificação da nota da avaliação.

| Classificação Individual | Intervalo | Observações |
|---|---|---|
| *Low Performers* | $<$ 94% | Não atingiu as expectativas definidas. |
| *On Target* | 95% - 105% | Atingiu totalmente as expectativas que foram definidas. |
| *High Performers* | 106% - 120% | Superou claramente todas as expectativas definidas. |

Considerando a amostra dos colaboradores activos, 52% da mesma atinge totalmente as expectativas que foram definidas e, quase 36% supera as expectativas predefinidas (3.9a). É necessário ter em consideração que, por motivos externos a este estudo, nem todos os colaboradores foram alvo de uma classificação individual.




(a) Activos.

(b) Saídas Voluntárias.

Figura 3.9: Distribuição dos colaboradores pela avaliação de desempenho.

A partir da figura 3.9b, pode-se verificar que colaboradores do grupo *On Target*, são potenciais *High Performers*, correspondendo assim à grande maioria da amostra. No entanto, apesar do grupo correspondente à avaliação mais elevada não representar a grande maioria, traduz-se na perda de potencial para a empresa. É importante referir novamente, que um dos principais objectivos deste projecto é reter os melhores talentos da entidade empresarial e, consequentemente, diminuir a percentagem de *turnover* voluntário destes colaboradores.

A partir da tabela 3.4 pode-se perceber como é distribuída a amostra consoante a avaliação de desempenho. A diferença de *performance* entre os colaboradores activos e os que saíram voluntariamente é desprezável.

Tabela 3.4: Características amostrais da variável avaliação de desempenho, consoante a amostra em estudo.

| Medida | Mínimo | Máximo | $\mathcal{Q}_{\frac{1}{4}}$ | $\mathcal{Q}_{\frac{1}{2}}$ | $\mathcal{Q}_{\frac{3}{4}}$ | Média | Desvio Padrão |
|---|---|---|---|---|---|---|---|
| Activos | 48 | 120 | 100 | 104 | 107 | 103,55 | 6,81 |
| Saídas Voluntárias | 47 | 117 | 100 | 104 | 107 | 103,12 | 8,13 |

Apesar das características amostrais das duas amostras serem praticamente iguais, é interessante perceber se, também em relação à existência de candidatos a *outlier*, esta igualdade se verifica.

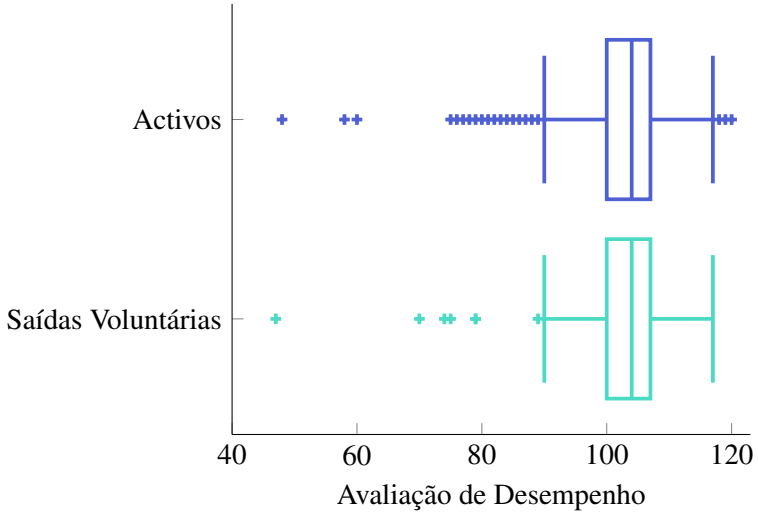



Figura 3.10: *Boxplots* paralelos da variável avaliação de desempenho dos colaboradores (Activos e Saídas Voluntárias).

Como se pode verificar através da figura 3.10, o comportamento face aos candidatos a *outlier* não é idêntico em ambas as amostras. Relativamente à amostra dos colaboradores activos, está-se perante 21 candidatos a *outlier* (48, 58, 60, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 118, 119 e 120), que corresponde a 48 indivíduos. Já na amostra que diz respeito às saídas voluntárias, pode-se verificar que existem apenas seis candidatos a *outlier* (47, 70, 74, 75, 79 e 89), o que corresponde a sete indivíduos da amostra.

A diversidade dos colaboradores é aparente nestas representações, em que se verifica que qualquer variável pode ser responsável pela categorização do perfil do colaborador que pretende rescindir contrato de forma voluntária.

# 4. Aplicação

Neste capítulo são apresentados os resultados da aplicação dos procedimentos já resumidamente descritos nos capítulos 2.1 e 2.2, modelo de regressão logística e árvores de decisão, respectivamente. No entanto, antes de iniciar o processo de cada metodologia é necessário definir estratégias de modelação. De seguida, serão apresentados os resultados obtidos de cada predição, consoante a metodologia utilizada.

O *software* utilizado para o cálculo da previsão de saída dos colaboradores de forma voluntária foi o *R Studio* ®. Para ser possível utilizar a ferramenta estatística foi necessário utilizar os seguintes *packages*: *Foreign; Lmtest; Faraway; Ltm; Oddsratio; Lmtest; pROC; OptimalCutpoints; Epi; Caret; Plyr; Rpart; Rattle; FSelector; ROCR* e *e1071*.

## 4.1 Estratégias de Modelação

De forma a manter a confidencialidade da empresa em estudo e, preservar todos os dados e informações utilizados neste projecto, define-se que, para a construção de cada modelo, são utilizadas $\mathcal{U}$ observações disponíveis.

Uma vez definida teoricamente no Capítulo 2 a metodologia, é necessário aplicá-la ao caso em estudo. Para tal, é necessário dividir a amostra em dois conjuntos distintos. Segundo Dangeti, 2017, a amostra é geralmente dividida aleatoriamente em 70% - 30% ou 80% - 20%, em conjunto de dados de treino e de teste, respectivamente. Ou seja (figura 4.1),

- *Training data*: representa o conjunto de observações usado para implementar a metodologia utilizada. Normalmente, são usadas 80% das observações. No entanto, para este estudo são utilizadas 70%. A escolha desta divisão provém da dimensão da amostra.
- *Test data*: corresponde ao conjunto de observações utilizado para testar o ajustamento do modelo. Assim, uma vez que se utiliza 70% da amostra para o ajuste do modelo, as restantes 30% observações são usadas para testar o modelo.

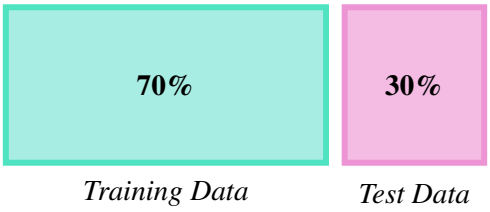



Figura 4.1: Estratégia de modelação estatística.

É necessário ter em consideração que, ao contrário do esperado, não é de elevada importância perceber o quão bem o método funciona no *training data*. Em vez disso, torna-se mais interessante para

o projecto a precisão das previsões que se obtêm quando se aplica o modelo no *test data* (Kassambara, 2018).

Por decisão da empresa, de forma a facilitar a interpretação dos modelos, foi definida uma nova categorização de duas variáveis, particularmente, a avaliação de desempenho e o estado civil. Portanto, a avaliação de desempenho é composta apenas por duas classes distintas, os *high performers* e os colaboradores cuja nota da avaliação individual é inferior a 106% ou que não possuem avaliação de desempenho. Em relação à variável estado civil, esta divide-se também em duas classes distintas: casados e não casados, que correspondem, respectivamente, aos indivíduos casados e aos colaboradores divorciados, viúvos ou solteiros. Por fim, relativamente à localização do estabelecimento de trabalho, uma vez que, a empresa é sediada na zona A, procurou-se encontrar apenas duas divisões geográficas, nomeadamente zona A e Outros territórios.

## 4.2 Diagnóstico e Conclusões do Modelo

Neste capítulo serão apresentados os resultados de cada modelo desenvolvido neste projecto e, por sua vez, os respectivos diagnósticos com o intuito de averiguar a capacidade discriminatória dos mesmos.

### 4.2.1 Regressão Logística

Um dos pressupostos do modelo de regressão logística é que as variáveis independentes utilizadas para a construção do modelo de regressão sejam não correlacionadas, isto é, a existência de multicolinearidade entre as variáveis pode traduzir-se num mau ajustamento do modelo. Pretende-se que em qualquer modelo não existam duas ou mais variáveis que traduzam a mesma informação. A título de exemplo, considere-se uma variável designada por IMC, Índice de Massa Corporal, que é calculado através do rácio entre o Peso (Kg) e Altura$^2$ (M). Um modelo composto por apenas estas três variáveis como independentes, não acrescenta valor ao mesmo, uma vez que, existe uma relação entre elas. Por outras palavras, é possível obter o IMC através das restantes variáveis.

Considere-se o modelo I composto pelas variáveis descritas no capítulo 3.3. De forma a averiguar a se, no modelo inicial, existe multicolinearidade, calcula-se o VIF para cada variável independente.

Tabela 4.1: *Variance Inflation Factors* das covariáveis do modelo I.

| **Variável** | **VIF** | **Variável** | **VIF** |
|---|---|---|---|
| Género | 1,07 | Distância Mercado Referência | 30,88 |
| Idade | 3,77 | Distância Mercado Geral | 67,02 |
| Antiguidade | 2,84 | Distância Mercado *Target* | 54,34 |
| Número de Filhos | 4,99 | % Aumento Salarial | 3,04 |
| Número de Dependentes | 5,20 | Número Promoções Funcionais | 2,82 |
| Local de Trabalho | 1,19 | Horizonte Temporal Funcional | 4,21 |
| Avaliação de Desempenho | 1,29 | Número Promoções Salariais | 6,39 |
| Estado Civil | 1,44 | Horizonte Temporal Salarial | 4,63 |
| Tipo de Contrato | 2,06 | Número Mobilidades | 2,20 |

| Variável | VIF | Variável | VIF |
|---|---|---|---|
| GO | 23,92 | Horizonte Temporal Mobilidades | 3,30 |
| Retribuição Salarial | 19,34 | | |

Como se pode verificar, através da tabela 4.1, é facilmente perceptível que a presença de multicolinearidade no modelo é bem evidente. Como existem valores superiores a 5, significa que existem variáveis dependentes entre si (Bicak et al., 2005).

As três variáveis que detêm os maiores valores de VIF são referentes à distância, em percentagem, do vencimento de cada colaborador face ao *target* salarial do seu GO, ao mercado referência e geral. Esta situação era de esperar, dado que, a empresa em estudo adopta uma estratégia salarial competitiva face ao mercado que a rodeia. Assim, conclui-se que a variável referente à distância do mercado geral está a ser explicada por outras variáveis já incluídas no modelo. Como tal, retira-se esta variável do modelo e calcula-se novamente o valor de VIF para cada covariável.

Tabela 4.2: *Variance Inflation Factors* das covariáveis do modelo I, da segunda iteração.

| Variável | VIF | Variável | VIF |
|---|---|---|---|
| Género | 1,07 | Retribuição Salarial | 19,80 |
| Idade | 3,78 | Distância Mercado Referência | 19,39 |
| Antiguidade | 2,81 | Distância Mercado *Target* | 33,06 |
| Número de Filhos | 5,07 | % Aumento Salarial | 3,03 |
| Número de Dependentes | 5,26 | Número Promoções Funcionais | 2,78 |
| Local de Trabalho | 1,19 | Horizonte Temporal Funcional | 4,17 |
| Avaliação de Desempenho | 1,29 | Número Promoções Salariais | 6,45 |
| Estado Civil | 1,44 | Horizonte Temporal Salarial | 4,67 |
| Tipo de Contrato | 2,09 | Número Mobilidades | 2,17 |
| GO | 22,67 | Horizonte Temporal Mobilidades | 3,25 |

Através da tabela 4.2, pode-se verificar que ainda se está perante a existência de multicolinearidade. Assim, pode-se concluir que ainda existe uma relação entre variáveis. Pretende-se repetir este processo de forma a que nenhum VIF associado a cada variável seja superior a cinco. Após quatro iterações deste processo, como se pode verificar através da tabela 4.3, não existe nenhum valor de VIF superior a 5.

É interessante perceber a forma como variam os valores de VIF para cada variável do modelo, à medida que é retirada uma variável. Uma vez que, a distância ao mercado de referência, mercado geral e ao *target* dependem do vencimento do colaborador, esperava-se que, ao final de todas as iterações, apenas uma das variáveis estivesse presente no modelo final. Como se pode verificar através da tabela, apenas a distância, em percentagem, do vencimento do colaborador ao valor mediano do mercado de referência permanece no modelo.

Tabela 4.3: *Variance Inflation Factors* das covariáveis do modelo I para cada iteração.

| **Variável** | $VIF_0$ | $VIF_1$ | $VIF_2$ | $VIF_3$ | $VIF_4$ |
|---|---|---|---|---|---|
| Género | 1,07 | 1,07 | 1,07 | 1,08 | 1,10 |
| Idade | 3,77 | 3,78 | 3,79 | 3,83 | 3,98 |
| Antiguidade | 2,84 | 2,81 | 2,78 | 2,79 | 2,76 |
| Número de Filhos | 4,99 | 5,07 | 5,02 | 4,93 | 4,69 |
| Número de Dependentes | 5,20 | 5,26 | 5,29 | 5,20 | 4,98 |
| Local de Trabalho | 1,19 | 1,19 | 1,15 | 1,16 | 1,18 |
| Avaliação de Desempenho | 1,29 | 1,29 | 1,28 | 1,17 | 1,27 |
| Estado Civil | 1,44 | 1,44 | 1,47 | 1,46 | 1,48 |
| Tipo de Contrato | 2,06 | 2,09 | 1,99 | 1,96 | 1,73 |
| GO | 23,92 | 22,67 | 5,12 | 4,94 | 1,84 |
| Retribuição Salarial | 19,34 | 19,80 | 5,83 | 5,82 | NA |
| Distância Mercado Referência | 30,88 | 19,39 | 2,50 | 2,53 | 1,32 |
| Distância Mercado Geral | 67,02 | NA | NA | NA | NA |
| Distância Mercado *Target* | 54,34 | 33,06 | NA | NA | NA |
| % Aumento Salarial | 3,04 | 3,03 | 2,99 | 2,04 | 2,06 |
| Número Promoções Funcionais | 2,82 | 2,78 | 2,63 | 2,40 | 2,42 |
| Horizonte Temporal Funcional | 4,21 | 4,17 | 4,05 | 3,53 | 3,62 |
| Número Promoções Salariais | 6,39 | 6,45 | 6,46 | NA | NA |
| Horizonte Temporal Salarial | 4,63 | 4,67 | 4,54 | 2,20 | 2,27 |
| Número Mobilidades | 2,20 | 2,17 | 2,15 | 2,12 | 2,17 |
| Horizonte Temporal Mobilidades | 3,30 | 3,25 | 3,26 | 3,22 | 3,24 |

Verificada a existência de multicolinearidade entre as variáveis do modelo, é necessário averiguar, através da *Deviance*, se, comparativamente a um modelo nulo, ao adicionar as variáveis a este modelo, a alteração no valor da *Deviance* é estatisticamente significativa. Sendo assim, pretende-se testar a qualidade de ajustamento do modelo. Uma vez que se pretende averiguar se não existem diferenças significativas entre os modelos e, dado que, o valor do *p-value* associado ao teste é aproximadamente zero, pode-se concluir que a qualquer nível de significância usual, rejeita-se a hipótese de que não existem diferenças significativas entre os dois modelos (tabela 4.4). Como tal, existe evidência para afirmar que os modelos diferem significativamente entre si.

Tabela 4.4: Teste de Razão de Verosimilhanças do modelo I.

| | | | | |
|---|---|---|---|---|
| **Modelo 1:** | Status $\sim 1$ | | | |
| **Modelo 2:** | Género + Estado Civil + Idade + Número de Filhos + Número de Dependentes + Antiguidade + Local de Trabalho + Tipo de Contrato + Avaliação de Desempenho + Número de Mobilidades + Número de Promoções Funcionais + Horizonte Temporal Mobilidades + Horizonte Temporal Promoções Funcionais + Horizonte Temporal Promoções Salarial + Distância Mercado Referência + GO + % Aumento Salarial | | | |
| | ***Resid. Dev*** | ***Df*** | ***Deviance*** | **Pr(> Chi)** |
| **Modelo 1:** | 439,24 | 15 | 90,189 | $\approx 0$ |
| **Modelo 2:** | 349,05 | | | |

Posto isto, procede-se ao método de selecção de variáveis, a fim de obter o melhor modelo. Assim, recorre-se ao método de selecção *Stepwise*. É necessário ter em consideração que, um modelo composto por um número maior de variáveis traduz-se numa melhor explicação da variável dependente e, por sua vez, mais parcimonioso. No entanto, esse modelo não será, obrigatoriamente, o melhor modelo sob o ponto de vista de predição.

O método de selecção progressiva através da inclusão e exclusão de variáveis, de acordo com o critério AIC, visa obter a escolha do melhor modelo. A decisão da adição de cada variável consiste na análise de testes $\mathcal{F}$ parciais, que são calculados para cada variável como se esta fosse adicionada pela primeira vez no modelo. Este método é utilizado de forma a obter a combinação ideal de variáveis independentes, visto que remove aquelas cuja importância no modelo é reduzida e, por sua vez, adiciona aquelas que mais contribuem para a variável dependente. O método finaliza quando não houver mais variáveis elegíveis para inclusão ou remoção no modelo (Marôco, 2018 e IBM, 2017).

Partindo do modelo composto pelas variáveis descritas na tabela 4.4 e recorrendo ao método de selecção de variáveis descrito anteriormente, obtém-se o modelo, designado por modelo II, composto por apenas sete variáveis: idade; antiguidade; género; distância ao mercado de referência; número de mobilidades; horizonte temporal referente às mobilidades e avaliação de desempenho (tabela 4.5).

Tabela 4.5: Sumário do modelo II.

| **Modelo** | ***Estimate*** | ***Std. Error*** | **z *Value*** | **Pr(>\| $z$ \|)** |
|---|---|---|---|---|
| *Intercept* | -0,089 | 1,242 | -0,071 | 0,943 |
| Idade | -0,165 | 0,038 | -4,350 | 1,36e-05 |
| Antiguidade | -0,077 | 0,046 | -1,665 | 0,096 |
| Género $_{\text{M}}$ | 0,574 | 0,321 | 1,786 | 0,074 |
| Mediana Referência | 1,170 | 0,751 | 1,557 | 0,119 |
| Número Mobilidades | 0,686 | 0,401 | 1,711 | 0,087 |
| Horizonte Temporal Mobilidades | 1,262 | 0,401 | 3,145 | 0,002 |
| Desempenho $_{\text{Outros}}$ | 1,277 | 0,346 | 3,687 | 0,000 |

Após a obtenção do modelo através do método de selecção da variáveis usado, é necessário aplicar novamente o teste de razão de verosimilhanças. Este teste é utilizado a fim de perceber se existem diferenças significativas entre o modelo obtido, através do método de selecção de variáveis, e o modelo antes do recurso a esta metodologia.

Tabela 4.6: Teste de Razão de Verosimilhanças do modelo II.

| | | | | |
|---|---|---|---|---|
| **Modelo 1:** | Status $\sim$ + Antiguidade + Género + | | | |
| | Distância Mediana de Referência + Número de Mobilidades + | | | |
| | Horizonte Temporal Mobilidades + Avaliação de Desempenho | | | |
| **Modelo 2:** | Status $\sim$ Género + Estado Civil + Idade + Número de Filhos + Número de Dependentes + | | | |
| | Antiguidade + Local de Trabalho + Tipo de Contrato + Avaliação de Desempenho + | | | |
| | Número de Mobilidades + Número de Promoções Funcionais + | | | |
| | Horizonte Temporal Mobilidades + Horizonte Temporal Promoções Funcionais + | | | |
| | Horizonte Temporal Promoções Salarial + Distância Mercado Referência + | | | |
| | GO + % Aumento Salarial | | | |
| | ***Resid. Dev*** | ***Df*** | ***Deviance*** | **Pr($>$ Chi)** |
| **Modelo 1:** | 349,05 | 10 | 5,181 | $\approx$ 0,8788 |
| **Modelo 2:** | 354,23 | | | |

Como se pode observar a partir da tabela 4.6, o *p-value* é superior aos níveis usuais de significância. Como se tratam de modelos encaixados, isto é, um dos modelos está incluído no outro, pode-se concluir que a adição das variáveis no modelo antes do uso da metodologia *stepwise* não é estatisticamente significativa. Como tal, escolhe-se o modelo mais parcimonioso, que corresponde ao modelo obtido através do método de selecção de variáveis.

Interpreta-se os valores dos efeitos de cada covariável do modelo para o contexto real do estudo através dos seus coeficientes associados, nomeadamente por meio da medida OR. No caso em estudo, esta medida quantifica a razão entre a probabilidade de sair de forma voluntária, definida como acontecimento de interesse, e a probabilidade de não sair de forma voluntária. Através da tabela 4.7 é possível verificar a estimativa dos coeficientes do modelo para cada covariável e o respectivo OR.

Tabela 4.7: Estimativas dos coeficientes do modelo II e respectivo OR.

| **Variável** | **Estimativa do Coeficiente** | **OR** |
|---|---|---|
| Idade | -0,165 | 0,848 |
| Antiguidade | -0,077 | 0,926 |
| Género $_{\text{M}}$ | 0,574 | 1,775 |
| Mediana Referência | 1,170 | 3,222 |
| Número Mobilidades | 0,686 | 1,986 |
| Horizonte Temporal Mobilidades | 1,262 | 3,532 |
| Desempenho $_{\text{Outros}}$ | 1,277 | 3,586 |

Sendo assim, pode-se interpretar que a possibilidade de uma saída voluntária de um indivíduo face a um indivíduo que permanece na empresa é derivada de diversos factores, nomeadamente:

- o valor da estimativa do coeficiente associado à variável idade é negativo e, consequentemente, o OR é inferior a 1. Como tal, por cada aumento unitário na idade, o OR diminui 0,152, ou seja, a chance de ocorrência de uma saída voluntária diminui 15,2%;
- por cada aumento unitário na antiguidade, o OR diminui 0,074, isto é, a chance de ocorrência de uma saída voluntária diminui $7,4\%$;
- o risco de um indivíduo masculino sair da empresa de forma voluntária é 0,775 vezes maior face a um indivíduo do género feminino;
- por cada aumento unitário na distância salarial face ao mercado referência, o OR aumenta 2,222;
- por cada aumento unitário no número de mobilidades, a razão entre a probabilidade de ocorrência de uma saída voluntária face a uma não ocorrência aumenta 0,986;
- por cada aumento unitário no horizonte temporal relativo às mobilidades, o OR aumenta 2,532
- o risco de saída de um colaborador, cuja avaliação individual é inferior a 106% ou que não possuiu um momento de avaliação, é 2,586 vezes maior face um indivíduo da classe *high performers*.

Definido o modelo, é necessário verificar a significância dos coeficientes do mesmo, através da aplicação do teste de *Wald*.

Tabela 4.8: Teste *Wald* do modelo II.

| ***Wald Test*** | | | |
|---|---|---|---|
| **Modelo 1:** | Status $\sim$ 1 | | |
| **Modelo 2:** | Status $\sim$ Idade + Antiguidade + Género + | | |
| | Distância Mediana de Referência + Número de Mobilidades + | | |
| | Horizonte Temporal Mobilidades + Avaliação de Desempenho | | |
| | ***Df*** | ***Chisq*** | ***Pr(> Chi)*** |
| | 7 | 67,398 | $\approx 0$ |

Como se pode observar a partir da tabela 4.8, o *p-value* é aproximadamente zero, sendo assim é possível concluir que as variáveis incluídas no modelo são estatisticamente significativas. Apesar de haver uma variável cujo *p-value* é superior aos níveis usuais de significância, nomeadamente a variável distância ao mercado de referência (tabela 4.5), é importante referir que, consoante o método de selecção de variáveis usado, esta variável deve ser incluída no modelo, uma vez que influencia a que o valor de AIC seja menor.

Quando se constrói um modelo de regressão, que contém várias variáveis independentes estatisticamente significativas, é comum tentar procurar saber qual é a variável mais importante. Facilmente se induz em erro ao estabelecer um grau de importância de uma variável, a partir do impacto de cada variável independente na variável resposta.

As estimativas dos coeficientes do modelo descrevem a relação de cada variável independente e a resposta. Ou seja, representa a alteração na variável resposta, dado um aumento de uma unidade na variável independente, caso se trate de uma variável quantitativa. Consequentemente, é fácil pensar que as variáveis com estimativas dos coeficientes maiores são mais importantes para o modelo, uma vez que,

Tabela 4.9: Importância de cada variável do modelo II.

| Variável | Importância |
|---|---|
| Idade | 4,350 |
| Avaliação de Desempenho | 3,687 |
| Horizonte Temporal Mobilidades | 3,145 |
| Género | 1,786 |
| Número de Mobilidades | 1,711 |
| Antiguidade | 1,665 |
| Distância Mercado Referência | 1,557 |

representam uma alteração maior na variável resposta. No entanto, as unidades variam entre os diferentes tipos de variáveis, o que torna impossível compará-las directamente. (Editor, 2016).

Através da tabela 4.9, pode-se verificar que a variável idade é aquela que corresponde à mais importante do modelo. No entanto, é também possível concluir que a variável referente à distância ao mercado referência é a que possui a menor importância. No ambiente empresarial, nem sempre as variáveis que, estatisticamente são consideradas mais importantes, correspondem às variáveis mais atractivas do ponto de vista do negócio (Editor, 2016).

Um dos pressupostos para o modelo de regressão logística é que os resíduos não apresentem padrão definido e que 95% dos resíduos estejam no intervalo $\left[-2;\ \ 2\right]$, uma vez que resíduos elevados (em valor absoluto) são o resultado de maus ajustamentos (Portugal, 2013). Note-se que um resíduo deve exprimir a discrepância entre o valor observado e o valor ajustado pelo modelo. Neste caso, tem-se que cerca de $97,45\%$ dos resíduos cumprem este requisito (figura 4.2).

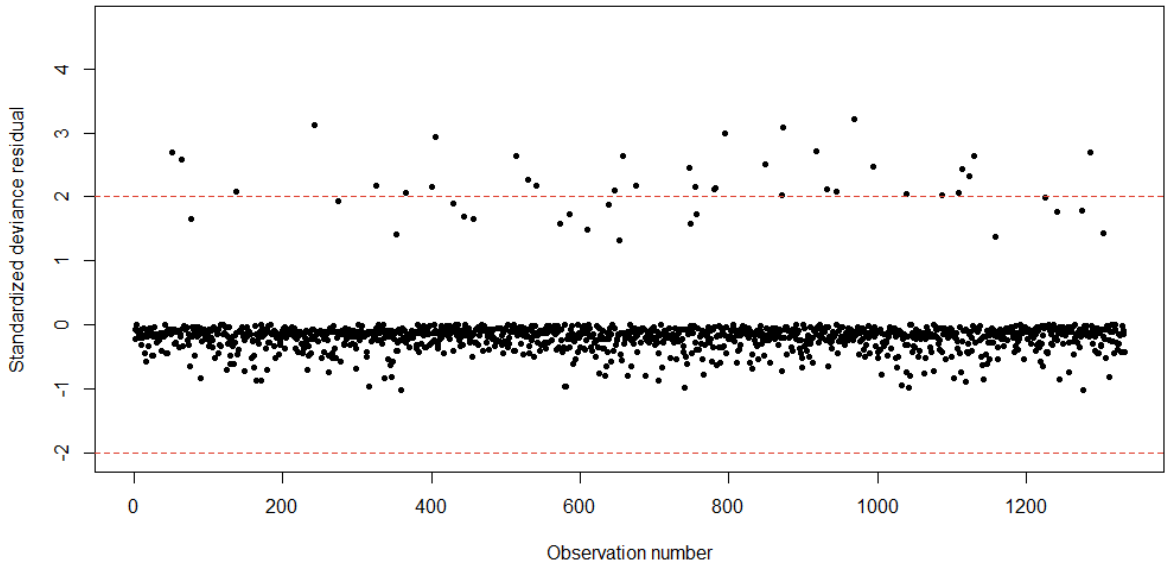



Figura 4.2: Resíduos padronizados do modelo II.

Uma observação influente é tal que, se a modificação ou a exclusão do modelo, produz alterações significativas nas estimativas dos parâmetros do modelo. Uma forma de averiguar a existência de observações influentes é através da análise da distância de *Cook*, como função da *leverage*, apresentada na figura 4.3. As observações identificadas na figura são as que apresentam maior distância de *Cook* e, consequentemente, são classificadas como potencialmente discordantes.

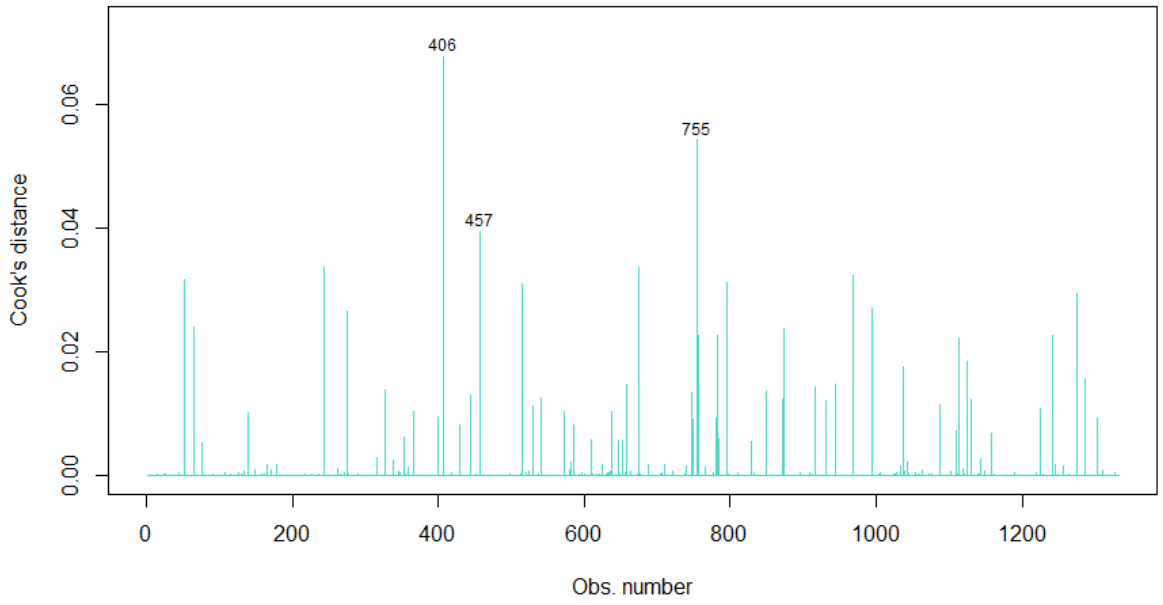



Figura 4.3: Distância de Cook do modelo II.

No entanto, estas observações possuem uma distância inferior a um não sendo, portanto, suficientemente elevada para considerar retirar as respectivas observações da amostra. Consequentemente, através da análise gráfica, pode-se concluir que não se está perante a existência de observações influentes no modelo.

Uma vez analisados os resíduos do modelo, está-se perante as condições necessárias para averiguar o ajuste do modelo aos dados em questão.

Como já foi referido anteriormente no capítulo 2.1.1.4, através da construção e análise da curva ROC e da construção da matriz de confusão é possível avaliar o ajustamento do modelo obtido. A curva ROC para o modelo em análise é apresentada na figura 4.4. Uma vez construída esta representação gráfica, é possível encontrar o *cut-off* óptimo. Este ponto traduz-se no melhor compromisso entre a taxa de falsos positivos e a taxa de verdadeiros positivos.

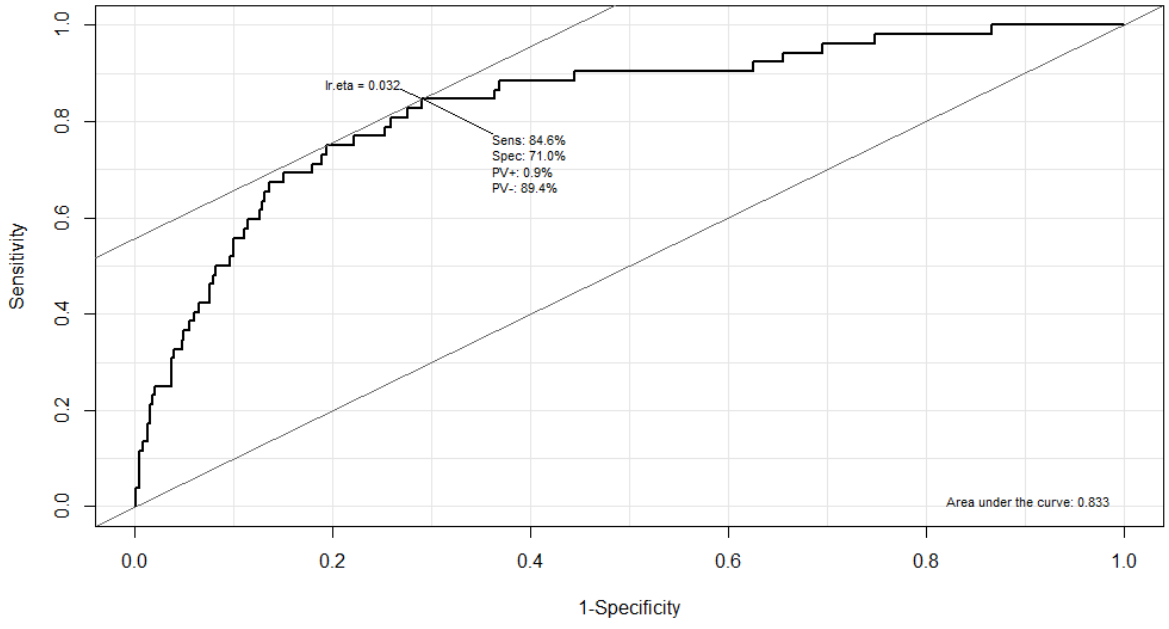



Figura 4.4: Curva ROC do modelo II.

Neste caso, o valor do *cut-off* é de 0,0318 e, consequentemente, a sensibilidade e a especificidade do modelo são $84,62\%$ e $71,02\%$, respectivamente. Ou seja, o modelo consegue prever correctamente cerca de 85% dos valores positivos. E, por sua vez, cerca de 71% das observações negativas são classificadas de forma correcta. A partir da figura 4.4 pode-se verificar que a AUC é cerca de $83,3\%$ e, segundo a tabela 2.2, pode-se afirmar que se está perante uma discriminação excelente. Posto isto, pode-se concluir que o modelo explica $83,3\%$ da amostra.

Dada a estratégia de modelação definida no capítulo 4.1, é necessário testar este modelo num conjunto de dados de teste, *test data*. É importante referir que este conjunto de observações nunca influenciou a construção do modelo tornando assim possível avaliar a capacidade discriminatória do modelo. Posto isto, aplica-se o modelo a apenas 30% das $\mathcal{U}$ observações e obtém-se a respectiva matriz de confusão (tabela 4.10).

Tabela 4.10: Matriz de confusão do modelo II.

| | **Observados** | |
|---|---|---|
| **Previstos** | 1 | 0 |
| 1 | 18 | 158 |
| 0 | 5 | 390 |

A partir dos valores da matriz de confusão é possível calcular diversas medidas que avaliam a qualidade do modelo em questão.

Tabela 4.11: Medidas de avaliação da qualidade do modelo II.

| **Medidas** | |
|---|---|
| Sensibilidade | $78,26\%$ |
| Especificidade | $71,17\%$ |
| *Accuracy* | $71,45\%$ |
| Eficiência | $74,71\%$ |

Através da tabela 4.11 pode-se verificar a validação do modelo e, por sua vez, quantificar o quão bom é o ajustamento do mesmo. O modelo desenvolvido aplicado ao *test data* apresenta resultados que não diferem do modelo aplicado ao *training data*. O modelo detém $78,26\%$ de capacidade de prever correctamente as observações que são classificadas como uma saída voluntária.

É também importante perceber o quão bem o modelo é capaz de prever as observações negativas, isto é, classificar um indivíduo que não irá sair da empresa de forma voluntária, dado que se trata de um indivíduo que permaneceu na empresa. Assim, o valor da especificidade é de $71,17\%$.

Pode-se ainda concluir que a proporção de predições correctas, que corresponde ao rácio entre o número de indivíduos classificados correctamente e o número total de observações, isto é, a *accuracy*, é cerca de 71%. Apesar deste valor ser consideravelmente alto, não se deve basear apenas nesta medida para medir a capacidade discriminatória do modelo, como já foi referido anteriormente no capítulo 2.1.1.4.

Por fim, calcula-se o valor da eficiência, de forma a contabilizar a média aritmética entre a sensibilidade e a especificidade. Quanto mais próximo este valor estiver de 100% maior é a precisão do modelo. Uma vez que o valor da eficiência é cerca de 75%, pode-se concluir que o modelo detém uma boa capacidade de previsão.

Uma vez que se está perante um número consideravelmente grande de variáveis, torna-se interessante perceber a relação que existe entre elas. Da mesma forma, aquando da criação de um potencial modelo, pretende-se estabelecer a melhor relação existente entre a variável dependente e as restantes variáveis independentes. Por exemplo, considere-se dois modelos iniciais, MOD1 e MOD2, compostos pelas seguintes variáveis: A, B, C e D, e A, B, E e F, respectivamente.

Como já foi demonstrado anteriormente, é efectuada uma análise exaustiva a fim de obter o modelo final que, por sua vez, seja capaz de predizer correctamente o maior número de observações. No entanto, o percurso necessário até à obtenção do modelo final resume-se em procurar relações entre as variáveis, interpretar as estimativas dos coeficientes das variáveis, analisar a influência de observações no modelo e, por fim, avaliar a qualidade do ajustamento do modelo.

Consoante as variáveis a incluir no modelo inicial, o comportamento do mesmo ao longo de cada etapa difere. Apesar dos modelos MOD1 e MOD2 terem apenas 2 variáveis em comum, o modelo final alcançado poderá ser ou não idêntico. Caso o modelo obtido não seja igual, podem-se obter conclusões sobre a qualidade de ajustamento do modelo bastante distintas. Desta forma, procura-se estabelecer a melhor combinação entre variáveis. É difícil perceber, *a priori*, quais são as variáveis que visam a obtenção de um potencial modelo. Como tal, o processo de encontrar o melhor modelo de previsão de saída voluntária dos colaboradores, consoante a escolha inicial das variáveis, torna-se bastante exaustivo dada a quantidade de variáveis existentes neste projecto.

Por isso, foi criado um algoritmo que permitisse criar o número máximo de modelos, partindo de cada combinação existente entre todas as variáveis. Por exemplo, considere-se a empresa Takk que contém apenas observações de quatro variáveis distintas, nomeadamente, o *status* do colaborador, a idade, o género e a antiguidade. O algoritmo calcula primeiramente o número de combinações existentes entre as três variáveis, visto que uma delas é, obrigatoriamente, a variável resposta. Sendo assim, estabelece combinações uma a uma, duas a duas e três a três. Perante este exemplo, o algoritmo cria os seguintes modelos:

- Status $\sim$ Idade
- Status $\sim$ Género
- Status $\sim$ Antiguidade
- Status $\sim$ Idade + Género
- Status $\sim$ Idade + Antiguidade
- Status $\sim$ Género + Antiguidade
- Status $\sim$ Idade + Género + Antiguidade

Para cada modelo estabelecido, o algoritmo verifica as condições necessárias para criar um modelo de regressão logística. Por fim, o algoritmo cria um ficheiro *Excel* com os resultados obtidos das medidas que permitem avaliar a qualidade do modelo, nomeadamente, a eficiência, a *accuracy*, a sensibilidade, a especificidade, a AUC, o *cut-off*, os VP, os VN, os FP e os FN. Não foi possível executar este algoritmo devido à incapacidade da máquina utilizada e das inúmeras horas necessárias para o efectuar. O *script* utilizado para este algoritmo pode ser consultado no apêndice B.

### 4.2.2 Árvores de Decisão

Conforme referido anteriormente, pretende-se aplicar um modelo de classificação que permite predizer a saída voluntária de um colaborador da empresa, com base nas suas características, a partir das variáveis apresentadas no capítulo 3.3.

De acordo com o princípio fundamental da ciência, conhecido como *Occam's razor*, quando se procura uma explicação para qualquer fenómeno, deve ser realizado o menor número possível de suposições e eliminar aquelas que não fazem diferenças na previsão observada das hipóteses explicativas (Rokach e Maimon, 2015).

A escolha do melhor modelo será obtida através da maximização da área abaixo da curva ROC, isto é, AUC. Este indicador de desempenho é muito mais imparcial do que a *accuracy* do modelo. Com esta medida de avaliação de desempenho do modelo, a *accuracy*, pode-se atingir valores muito baixos de verdadeiros positivos e um número alto de verdadeiros negativos. Do ponto de vista do negócio da empresa, esta medida não fornece muitas informações, pois o objectivo é identificar o maior número de verdadeiros valores positivos e, por sua vez, aumentar a retenção dos colaboradores (Graça et al., 2017).

Como já foi referido anteriormente, o ganho de informação corresponde à redução esperada na entropia causada pela nova divisão dos dados, de acordo com um determinado atributo. Sendo assim, a primeira divisão da árvore de decisão traduz-se na variável que contém o maior ganho de informação.

Tabela 4.12: Ganho de Informação.

| **Variável** | **Ganho de Informação** |
|---|---|
| Idade | 0,019189 |
| Antiguidade | 0,017727 |
| Retribuição | 0,008930 |
| Número de Dependentes | 0,007893 |
| Número de Filhos | 0,007365 |
| Tipo de Contrato | 0,005979 |
| Estado Civil | 0,005692 |
| Local | 0,002599 |
| Avaliação de Desempenho | 0,001164 |
| Género | 0,000406 |
| Número de Mobilidades | 0 |
| Número de Promoções Funcionais | 0 |
| Horizonte Temporal Mobilidades | 0 |
| Horizonte Temporal Promoções Funcionais | 0 |
| Número de Promoções Salariais | 0 |
| Horizonte Temporal Promoções Salariais | 0 |
| Distância à Mediana de Referência | 0 |
| Distância à Mediana de Geral | 0 |

| Variável | Ganho de Informação |
|---|---|
| Distância ao *Target* | 0 |
| GO | 0 |
| % Aumento Salarial | 0 |

A partir da tabela 4.12, pode-se verificar que existem variáveis que não trazem um ganho de informação. No entanto, também se pode constatar que a variável idade é a que possui um maior ganho de informação e, consequentemente, corresponde ao primeiro nó de divisão da árvore de decisão. As divisões seguintes são obtidas através da proporção amostral existente entre cada variável, face às saídas voluntárias e os colaboradores que se mantiveram na empresa. A figura 4.5 representa a primeira árvore de decisão para este modelo.

De forma a manter a confidencialidade dos dados da empresa, os valores associados às divisões de cada nó são meramente ilustrativos.

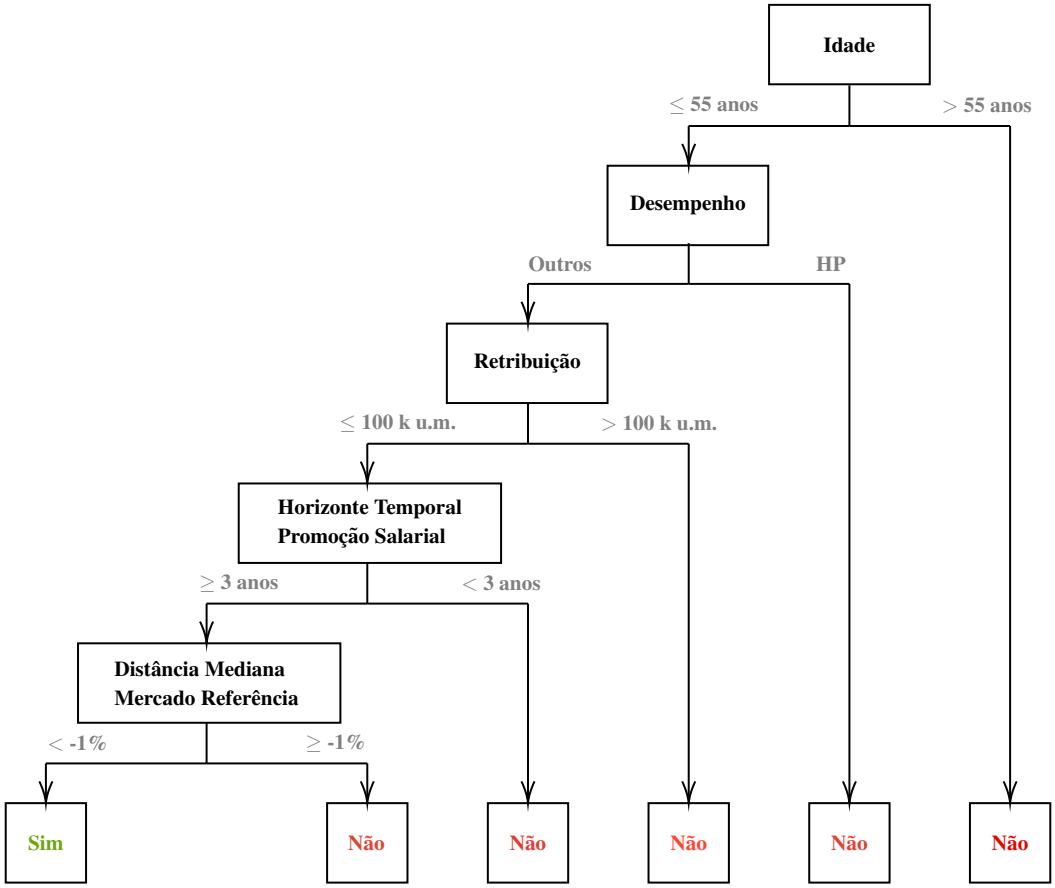



Figura 4.5: Árvore de decisão do modelo.

A árvore da figura 4.5 corresponde à representação da divisão estabelecida da amostra, de forma a estabelecer o padrão dos colaboradores que saíram de forma voluntária. Como se pode verificar, é necessário efectuar cinco divisões da amostra para encontrar a razão pela qual os colaboradores efectuam a rescisão do contrato de forma voluntária. Como já foi referido anteriormente, a variável idade é a que permite efectuar a primeira divisão da amostra. Como tal, caso os colaboradores tenham uma idade

superior a 55 anos, segundo o algoritmo utilizado, não existem evidências para afirmar que o colaborador irá sair da empresa, de forma voluntária.

Por outro lado, caso os colaboradores tenham as seguintes características, nomeadamente:

- idade inferior ou igual a 55 anos;
- a avaliação de desempenho ser diferente de *high performers*;
- a retribuição anual ser inferior ou igual a 100 mil unidades monetárias;
- pelo menos há três anos não foi sujeito a uma promoção salarial;
- a distância face à mediana do mercado de referência é inferior a menos um ponto percentual,

são considerados como um grupo potencial a rescindir contrato de forma voluntária.

A aprendizagem e validação deste modelo foi realizada através da metodologia mencionada no capítulo 4.1. No entanto, os 70% da amostra em estudo são também utilizados para validar o modelo, como mencionado no capítulo 2.2.2.

Neste projecto, o método de validação cruzada baseia-se na validação de *k-fold*, com $k = 5$. O valor de $k$, como já foi referido anteriormente, pode variar entre 10 e 5. No entanto, se o método de validação cruzada for o *leave-one-out*, tem-se que, $k$ corresponde a 70% das $\mathcal{U}$ observações. Este método de validação não é utilizado, uma vez que se trata de um processo bastante exaustivo.

A título de exemplo, a empresa Vennskap é constituída por 5 colaboradores, nomeadamente, A; B; C; D e E. Se o método de validação cruzada for o *leave-one-out* tem-se o seguinte processo (tabela 4.13):

Tabela 4.13: Esquema do processo de uma validação cruzada de k-*folds*, onde k = n = 5.

| **Membros da amostra** | **Treino** | **Teste** |
|---|---|---|
| Iteração 1 | A, B, C, D | E |
| Iteração 2 | A, B, C, E | D |
| Iteração 3 | A, B, D, E | C |
| Iteração 4 | A, C, D, E | B |
| Iteração 5 | B, C, D, E | A |

O processo de validação cruzada repete-se assim cinco vezes, de forma a que, a escolha da amostra seja o mais aleatória possível. Para além disso, foi definido um *minsplit* de 15, ou seja, o número mínimo de observações que deve existir num nó para que exista uma nova divisão dos dados, isto é, um nó filho.

Através da validação cruzada é possível obter o melhor parâmetro de complexidade, $\lambda$, com o intuito de determinar a extensão do *pruning*. Na aplicação das árvores de decisão a grandes volumes de dados para processamento, o *pruning* é uma abordagem eficaz na optimização de tempo de processamento, contudo existe uma perda de precisão do algoritmo. Dependendo das áreas de estudo esta perda de precisão pode ou não ser relevante (Ferreira, 2013).

O parâmetro de complexidade é usado para controlar o tamanho da árvore e para seleccionar o tamanho ideal. Se o custo associado à adição de outra variável à árvore de decisão no nó actual estiver acima do valor de $\lambda$, a construção da árvore não continuará (Siddhant, 2015).

A partir da tabela 4.14 é possível verificar a relação existente entre o parâmetro de complexidade face a AUC, a sensibilidade e a especificidade.

Tabela 4.14: Relação existente entre o parâmetro de complexidade e as medidas de avaliação de desempenho.

| $\lambda$ | **AUC** | **Sensibilidade** | **Especificidade** | $\lambda$ | **AUC** | **Sensibilidade** | **Especificidade** |
|---|---|---|---|---|---|---|---|
| 0 | 0,589 | 0,02 | 0,991 | 0,013 | 0,589 | 0,02 | 0,991 |
| 0,001 | 0,589 | 0,02 | 0,991 | 0,014 | 0,589 | 0,02 | 0,991 |
| 0,002 | 0,589 | 0,02 | 0,991 | 0,015 | 0,542 | 0,02 | 0,992 |
| 0,003 | 0,589 | 0,02 | 0,991 | 0,016 | 0,542 | 0,02 | 0,992 |
| 0,004 | 0,589 | 0,02 | 0,991 | 0,017 | 0,542 | 0,02 | 0,992 |
| 0,005 | 0,589 | 0,02 | 0,991 | 0,018 | 0,542 | 0,02 | 0,992 |
| 0,006 | 0,589 | 0,02 | 0,991 | 0,019 | 0,542 | 0,02 | 0,992 |
| 0,007 | 0,589 | 0,02 | 0,991 | 0,02 | 0,542 | 0,02 | 0,992 |
| 0,008 | 0,589 | 0,02 | 0,991 | 0,021 | 0,542 | 0,02 | 0,992 |
| 0,009 | 0,589 | 0,02 | 0,991 | 0,022 | 0,542 | 0,02 | 0,992 |
| 0,01 | 0,589 | 0,02 | 0,991 | 0,023 | 0,542 | 0,02 | 0,992 |
| 0,011 | 0,589 | 0,02 | 0,991 | 0,024 | 0,5 | 0 | 1 |
| 0,012 | 0,589 | 0,02 | 0,991 | 0,025 | 0,5 | 0 | 1 |

Como se pode verificar, quando o $\lambda$ é igual a 0.014 tem-se que o valor da área abaixo da curva ROC é de 0,589, ou seja, 58,9% da amostra é explicada pelo modelo. Uma vez que a métrica escolhida para a escolha do modelo é baseada no AUC, então o valor do parâmetro de complexidade deverá ser tal que maximize a área abaixo da curva de ROC. Deste modo, tem-se que $\lambda$ igual a 0,014 e, por sua vez, será utilizado para efectuar o *pruning* da árvore de decisão. Como tal, o valor da Sensibilidade e da Especificidade na amostra de teste são, respectivamente, 2% e 99,1%.

Dado que o valor do AUC não se altera para $\forall \lambda, \lambda \in [0; 0,014]$, e como se pretende a maximização da área abaixo da curva ROC, então a utilização do método de *pruning* irá conduzir à árvore inicial, apresentada na figura 4.5.

Posto isto, está-se nas condições necessárias para testar o modelo na amostra de teste. Ao aplicar este modelo, o valor de AUC da amostra de teste obtido é de 66,85% e, como tal, pode-se afirmar que este modelo explica cerca de 67% da amostra de teste.

Por fim, pretende-se saber a capacidade discriminatória do mesmo e, como já foi mencionado nos capítulos 2.1.1.4 e 2.2.3 existem diversas métricas que permitem averiguar o ajustamento do modelo.

Tabela 4.15: Matriz de confusão do modelo de árvores de decisão.

| | **Observados** | |
|---|---|---|
| **Previstos** | 1 | 0 |
| 1 | 3 | 3 |
| 0 | 20 | 545 |

*A priori*, consegue-se perceber que a capacidade do modelo prever os verdadeiros positivos é bastante baixa, comparando com o modelo de regressão logística (tabela 4.10). O modelo obtido através da

metodologia árvores de decisão detém uma maior capacidade de prever os verdadeiros negativos.

Tabela 4.16: Medidas de avaliação da qualidade do modelo de árvores de decisão.

| **Medidas** | |
|---|---|
| Sensibilidade | 13,04% |
| Especificidade | 99,45% |
| *Accuracy* | 95,97% |
| Eficiência | 56,25% |
| Precisão | 50,00% |
| F | 20,69% |
| Prevalência | 4,03% |
| Taxa de Falsos Positivos | 0,55% |
| Taxa de Falsos Negativos | 89,96% |

De acordo com a tabela 4.16, consegue-se perceber através do valor da sensibilidade que o modelo é apenas capaz de prever cerca de 13% das observações positivas, ou seja, classificar um indivíduo como positivo e, que de facto, é positivo. Pode-se ainda concluir que a proporção de predições correctas é de 95,97%. Este valor é bastante elevado, no entanto, não traduz uma boa capacidade discriminatória do modelo, uma vez que, está influenciado pela capacidade do modelo prever correctamente as observações negativas (99,45%).

Uma boa medida de avaliação do modelo é aquela que é capaz de analisar o modelo como um todo, incluindo a capacidade de prever correctamente tanto as observações negativas como as positivas. Como tal, tem-se que o valor da eficiência é 56,25%.

Relativamente ao número de valores ajustados e, por sua vez, classificados como positivos, apenas 50% corresponde à proporção de verdadeiros positivos. Como a medida F depende do valor da sensibilidade e do da precisão e, dado que os respectivos valores são considerados baixos, espera-se que o valor associado à medida F seja também baixo. Como se pode verificar, o valor desta medida é de 20,69%.

Apenas 4,03% da amostra de teste corresponde aos valores observados classificados como positivos. A capacidade do modelo prever erradamente uma observação positiva e negativa é, respectivamente, 0,55% e 89,96%.

Uma vez que, através da primeira metodologia utilizada, a regressão logística, o modelo obtém uma maior eficiência e, por sua vez, detém maior capacidade em detectar os verdadeiros positivos, define-se como o modelo a adoptar pela empresa em estudo.

A escolha do modelo final deve-se à melhor conjugação dos valores obtidos para as medidas de avaliação da qualidade de ajustamento do modelo. Caso o propósito da empresa fosse encontrar o maior número de verdadeiros negativos, isto é, procurar saber quais são os potenciais colaboradores a permanecer na empresa, o modelo escolhido seria aquele que apresentasse um valor superior de especificidade, uma vez que se trataria da capacidade do modelo prever correctamente os verdadeiros negativos.

Para além deste trabalho desenvolvido, foi tido em conta outra amostra da empresa em estudo, referente a outro horizonte temporal. Essa amostra é constituída por $\mathcal{T}$ observações para $\mathcal{N}$ variáveis. Apesar da capacidade discriminatória da metodologia das árvores de decisão ser inferior à da regressão logística, procurou-se verificar se, ao construir um novo modelo para outro horizonte temporal e composto por outras variáveis, a qualidade do ajustamento do modelo seria superior à obtida. Ou seja, nem

sempre as metodologias se comportam da mesma forma ao longo do horizonte temporal utilizado, devido às variações existentes nas variáveis, à adição de outras variáveis e à adequabilidade do objectivo do projecto.

Tabela 4.17: Medidas de avaliação da qualidade do modelo adicional de árvores de decisão.

| **Medidas** | |
|---|---|
| Sensibilidade | 25,00% |
| Especificidade | 99,45% |
| *Accuracy* | 96,18% |
| Eficiência | 62,23% |
| Precisão | 67,74% |
| F | 36,52% |
| Prevalência | 4,39% |
| Taxa de Falsos Positivos | 0,55% |
| Taxa de Falsos Negativos | 75,00% |

Como se pode verificar através da tabela 4.17, a capacidade do modelo prever correctamente as observações positivas é de 25%. É possível verificar que, relativamente às observações classificadas como negativas, ou seja, os colaboradores que permanecem na empresa, o modelo prevê correctamente cerca de 99%. Uma vez que o valor da sensibilidade é superior ao do modelo da figura 4.5 e o valor da especificidade é idêntico em ambos os modelos, sabe-se que o valor da eficiência é, consequentemente, superior. Em relação ao número de valores ajustados e classificados como positivos, cerca de 68% correspondem aos verdadeiros positivos. É também possível verificar que, existe um aumento face ao valor da medida F, uma vez que tanto os valores da sensibilidade como da precisão aumentaram. Apenas 4,39% da amostra corresponde aos valores observados como positivos, ou seja, que rescindiram contrato de forma voluntária. A capacidade do modelo prever erradamente uma observação positiva e negativa é, respectivamente, 0,55% e 75%.

# 5. Conclusão e Trabalho Futuro

Nas últimas décadas é possível verificar o elevado e rápido crescimento tecnológico. Este avanço permitiu adquirir cada vez mais conhecimento técnico e analítico. Actualmente, é difícil encontrar organizações que não tenham por base o uso da tecnologia, com intuito de gerar recomendações accionáveis, no ponto de vista do negócio. Contudo, para conseguir gerar *insights* é necessário possuir competências analíticas, que permitem extrair conhecimento útil dos dados.

Dada a competitividade do mercado empresarial e do desenvolvimento de áreas de interesse para o negócio, por exemplo, o *people analytics*, é cada vez mais importante as empresas encontrarem modelos sofisticados para a tomada de decisão. Ao mesmo tempo é essencial uma limpeza dos dados para garantir a qualidade e a consistência destes, por forma a assegurar resultados sólidos e confiáveis.

Este trabalho procurou dar resposta a uma realidade presente nas empresas a nível mundial. O tema da elevada rotatividade dos colaboradores no mercado empresarial é cada vez mais tido em conta pelas organizações. É de elevado interesse o desenvolvimento de técnicas capazes de detectar e antecipar o possível *flight* dos colaboradores. Assim, traduz-se no estudo da caracterização de potenciais colaboradores a rescindirem o contracto de forma voluntária. Como tal, consoante o perfil de cada membro da organização, é necessário analisar a forma como a saída voluntária se expressa na relação existente entre o colaborador e a empresa. Um modelo de retenção para o ramo empresarial é uma abordagem recente no mercado português. Este foi criado com o propósito de identificar e quantificar os colaboradores que pretendem rescindir o contrato empresarial, deste modo, está-se perante um modelo de classificação.

Assim, numa primeira fase, foi efectuado um enquadramento geral do problema em análise. Em seguida foi feita uma descrição de como é que a utilização de métodos analíticos preditivos podem potenciar o crescimento empresarial do ponto de vista de gestão de pessoas. Por sua vez, é explicado quais são as etapas necessárias para a obtenção de um ecossistema analítico. Por fim, são apresentados casos de sucesso da utilização de políticas de HRA.

Numa segunda fase apresentaram-se alguns fundamentos teóricos sobre a metodologia utilizada, bem como os seus pressupostos. Seguidamente foi descrito a identificação de cada variável, seguindo-se de uma breve análise de dados, tendo em conta a confidencialidade dos mesmos. Por fim, procedeu-se à definição da estratégica de modelação utilizada, a fim de obter as melhores conclusões dos modelos criados, consoante as metodologias definidas anteriormente.

Os resultados sumários deste estudo, através da metodologia da regressão logística, sugerem que as variáveis incluídas no modelo não possuem o mesmo nível de importância, sendo a variável idade aquela que possui uma maior importância no modelo. Posto isto, é possível predizer correctamente as observações positivas e negativas em, respectivamente, 78,26% e 71,17%. Consequentemente, pode-se concluir que o modelo detém uma boa capacidade de previsão, uma vez que, o valor da eficiência é de 75%.

Relativamente ao modelo obtido através do uso das árvores de decisão, a variável idade é a que contribui para o maior ganho de informação do modelo, o que corrobora a situação já existente no

modelo de regressão logística. No entanto, este modelo não detém uma boa capacidade de previsão, aquando comparado com o obtido através da primeira metodologia.

Deste modo, o primeiro modelo obtém uma maior eficiência e, por sua vez, detém uma maior capacidade em detectar os verdadeiros positivos. Consequentemente, define-se como o modelo a adoptar pela empresa em estudo.

Não obstante, foi feita a construção de um modelo alternativo aplicado a outro horizonte temporal, composto pelas mesmas variáveis. O objectivo deste modelo alternativo vai ao encontro do modelo desenvolvido neste projecto, através do uso das árvores de decisão e de um conjunto de dados disponibilizados. De certa forma, não se pode comparar os resultados obtidos desde modelo com os dois primeiros modelos, uma vez que não se trata do mesmo horizonte temporal e da mesma amostra. Porém, o principal intuito desta modelação foi, de certa forma, perceber se uma vez que o modelo obtido pelo uso da segunda metodologia detém uma capacidade discriminatória inferior ao da regressão logística, esta metodologia aplicada a outro ambiente amostral poderia, ou não, obter resultados superiores. Apesar do valor da sensibilidade ter aumentado mas não ser o mais apelativo, pode-se verificar que a mesma metodologia pode ter desempenhos diferentes, consoante o meio ambiente de dados a que se aplica. Este processo torna-se interessante com o intuito de encontrar a metodologia que mais se ajusta aos dados ao longo do tempo, o que se traduz num possível trabalho futuro.

Há muito trabalho a desenvolver no que se refere à previsão de saída dos colaboradores da empresa. Por exemplo, a utilização de uma validação que reflicta a experiência que se pretende pôr em prática. Isto é, utilizar a amostra de dados no momento anterior ao ponto de validação, de modo a que o treino do modelo seja sempre com o passado, de forma a prever o futuro.

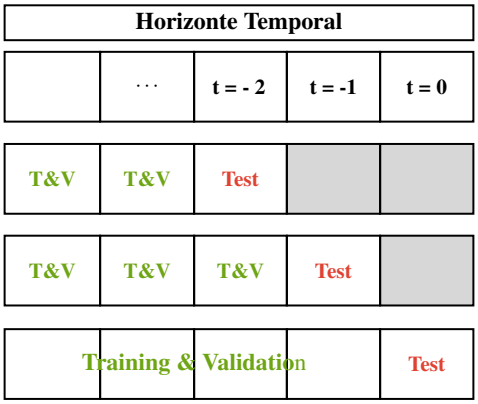



Figura 5.1: Predizer para o dia seguinte, utilizando os dados dos momentos temporais anteriores.

A partir da figura 5.1, pode-se verificar que ao longo do horizonte temporal são efectuados diversos treinos e validações de cada modelo. Este treino consecutivo tem como intuito aplicar o melhor modelo obtido de cada validação efectuada.

Para além do uso desta metodologia, um possível trabalho futuro caracteriza-se pelo uso de séries temporais. Isto permite obter informação mais detalhada de cada colaborador durante o seu percurso na empresa. Simplesmente captar a última observação de cada variável pode, ou não, conter toda a informação do colaborador desde a sua entrada na empresa. Assim, as variáveis criadas, através deste método, poderão permitir observar o percurso completo de um colaborador e, que por sua vez, será capaz de ser mais preciso para a obtenção de padrões de saídas voluntárias. Contudo, poderá não ser necessariamente verdade, mas é necessário que se esteja perante um conjunto de variáveis que não estagnam ao longo do horizonte temporal, não tendo em conta variáveis que se alteram obrigatoriamente ao longo do tempo, por exemplo, a idade e a antiguidade.

É necessário ter em conta que o modelo obtido foi um ajuste à realidade da empresa, pelo que pode variar ao longo dos anos. Sendo assim, aquando da revisão do mesmo, deve-se tentar incluir outras variáveis de modo a que seja possível explicar melhor o comportamento dos colaboradores. Seria também interessante abordar outras metodologias mais robustas, por exemplo *random forest*, *artificial neural network* e *support vector models*, com o intuito de, mais uma vez, comparar os modelos obtidos.

# Bibliografia


Agresti, Alan (2019). *An Introduction to Categorical Data Analysis*. Third Edit. Wiley.

Ahituv, Avner e Robert I. Lerman (2005). "Job turnover, wage rates, and marital stability: How are they related?" Em: *Review of Economics of the Household* 9, pp. 221–249.

Alpuim, Teresa (2018). *Apontamentos da Unidade Curricular de Modelos Lineares*.

Andreozzi, V. (2012). *Modelo Linear Generalizado*.

Antunes, Marília (2009). *CRM e Prospecção de Dados*.

Bermudez, Patrícia (2019). *Apontamentos da Unidade Curricular de Modelos Lineares Generalizados*.

Bersin, Josh (2014). "The datafication of HR". Em: *Deloitte Review issue 14*.

Bicak, Hasan et al. (2005). "Forecasting the Tourism Demand of North Cyprus". Em: *Journal of Hospitality & Leisure Marketing* 12, pp. 87–99.

Biggs, Josh (2019). *Real-World Ways to Reduce Employee Turnover*.

Blake, Irene A. (2020). *What Do You Do if You Live Too Far From Your Job?*

Bollen, Kenneth e Robert Jackman (1985). "Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases". Em: *Sociological Methods & Research* 13.4, pp. 510–542.

Bratko, Ivan e Marko Bohanec (1994). *Trading accuracy for simplicity in decision trees*. Springer.

Breiman, Leo et al. (1984). *Classification and Regression Trees*. Taylor & Francis Ltd.

Brynjolfsson, Erik et al. (2014). "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" Em: *The Datafication of HR*.

Chatzidimitriou, Kyriakos et al. (2018). *Practical Machine Learning in R*.

Cook, R. Dennis e Sanford Weisberg (1982). *Residuals and Influence in Regression*. Chapman e Hall.

Dangeti, Pratap (2017). *Statistics for Machine Learning*. Packt Publishing Ltd.

Deloitte, Bersin by (2013). "High-Impact Talent Analytics: Building a World-Class HR Measurement and Analytics Function". Em: *Whatworks Brief*.

Dobson, Annette J. e Adrian G. Barnett (2008). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC, Taylor & Francis Group.

Dostie, Benoit (2005). "Job Turnover and Returns to Seniority". Em: *Journal of Business & Economic Statistics* 23, pp. 192–199.

Duggan, Tara (2020). *The Effects of Salary on Job Retention*.

Economias (2016). *Os 8 tipos de rescisão de contrato de trabalho*.

Editor, Minitab Blog (2016). *The Minitab Blog*.

EliteDataScience (2019). *Overfitting in Machine Learning: What It Is and How to Prevent It*.

Empresa (2017). *HR Analytics, Leading to Excellence*. Rel. téc.

— (2018). *Relatório e Contas*. Rel. téc.

EURONEXT (2018). "Index Rule Book - PSI 20 Index, Version 18-01a". Em: *Live Euronext*.

Faraway, Julian (2009). *Texts in Statistical Science: Linear Models with R*. Chapman & Hall/CRC, Taylor & Francis Group.

Ferreira, Dário Miguel Ribeiro (2013). "Árvores de Decisão Aplicadas à Deteção de Formas Costeiras Através de Imagens IKONOS-2". Tese de mestrado.

Gomes, Bruno Miguel Viana (2011). "Previsão de Churn em Companhias de Seguros". Tese de mestrado. Universidade do Minho.

Graça, Carlos et al. (2017). "Big Data and Data Science for Business Analytics - Bank Marketing Project". Tese de mestrado.

Gregory, Ellen (2019). *The Age of Employment Turnover*.

IBM (2017). *IBM SPSS Statistics Brief Guide*.

Institute, HRO Today (2015). "Best Practices for Tying HR Metrics to Business Outcomes". Em: *Alexander Mann Solutions*.

IRH (2019). "O novo ciclo do talento nas organizações é o tema da RHconferência". Em: *infoRH*.

J.Faraway, Julian (2006). *Extending the Linear Model with R Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC, Taylor & Francis Group.

Jung, Tommy et al. (2014). *Predictive Analytics for Dummies*. John Wiley & Sons, Inc.

Kassambara, Alboukadel (2018). *Machine Learning Essentials: Pratical Guide in R*. CreateSpace Independent Publishing Platform.

Kutner, Michael H. et al. (2004). *Applied Linear Statistical Models*. Fifht Edit. Vol. 149. 1. McGraw-Hill/Irwin.

Lee, Tae Heon (2012). "Gender Differences in Voluntary Turnover: Still a Paradox?" Em: *International Business Research* 5.10, pp. 19–28.

Lewis, Gregory (2018). *How Hershey Used Data to Increase Retention Rates and Improve Workforce Planning*.

Madsen, Dag e Kåre Slåtten (2017). "The Rise of HR Analytics: A Preliminary Exploration". Em: *Business and Finance Proceedings*.

Marler, Janet e John Boudreau (2016). "An evidence-based review of HR Analytics". Em: *The International Journal of Human Resource Management*, pp. 1–24.

Marôco, João (2018). *Análise Estatística com o SPSS Statistics*. ReportNumber.

Menard, Scott W. (1995). *Applied Logistic Regression Analysis*. Sage Publications, Inc.

Montegomery, Douglas C. et al. (2012). *Introduction to Linear Regressional Analysis*. Fifth Edit. Wiley.

Myres, Raymond H. et al. (2010). *Generalized Linear Models with Applications in Engineering and the Sciences*. Second Edi. Wiley.

Notícias, Jornal de (2011). *Quem são os elementos do agregado familiar*.

Portugal, Marta Gonçalves Cruces Simão (2013). "Modelos Estatísticos para a Previsão de Inactividade de Pré-Pagos". Tese de mestrado.

República, Assembleia da (2019). *Diário da República Eletrónico*.

República, Diário da (2009). *Código do Trabalho - Lei n.º 7/2009*.

Rokach, Lior e Oded Maimon (2015). *Data Mining With Decision Trees - Theory and Applications - 2nd Edition*. 2nd Editio. World Scientific Publishing Co. Pte. Ltd.

Sammut, Claude e Geoggrey I. Webb (2017). *Encyclopedia of Machine Learning and Data Mining (2nd edition)*. Second Edi. Vol. 32. 7/8. Springer.

Scott, A. J. et al. (2013). *Applied Logistic Regression*. John Wiley & Sons, Inc.

Siddhant (2015). *Analytics Vidhya*.

Siegel, Eric (2013). *Predictive Analytics: The privacy pickle – Hewlett-Packard's prediction of employee behavior*.

Smyth, Padhraic et al. (1996). *Advances In Knowledge Discovery And Data Mining*. MIT Press Ltd.

SocialChorus (2019). *High Employee Turnover? – The Real Causes And Impact*.

Torrejano, Alexandre Pereira Martins Rafael (2018). "Desagregação de consumos de smart homes por tipologia". Tese de mestrado.

Turkman, M. Antônia Amaral e Giovani Loiola Silva (2000). *Modelos Lineares Generalizados: da teoria à prática.*

Westfall, Brian (2017). *How SMBs Can Begin to Assess Employee Flight Risk*.

# Apêndices

# A Algoritmo para obter os padrões de resposta

```
# CRIACAO DAS COMBINACOES POSSIVEIS ENTRE OS MOTIVOS!

for (VariavelIndicador in Indicadores) {
  if (VariavelIndicador == "Idade") {
    subbase = SubBaseIdade
    SubBaseExcel = SubBaseIdadeExcel
  }
  if (VariavelIndicador == "Antiguidade") {
    subbase = SubBaseAntiguidade
    SubBaseExcel = SubBaseAntiguidadeExcel
  }
  if (VariavelIndicador == "EscalaoDesemp") {
    subbase = SubBaseDesempenho
    SubBaseExcel = SubBaseDesempenhoExcel
  }
  if (VariavelIndicador == "GO") {
    subbase = SubBaseGO
    SubBaseExcel = SubBaseGoExcel
  }
  if (VariavelIndicador == "Cluster") {
    subbase = SubBaseCluster
    SubBaseExcel = SubBaseClusterExcel
  }

  for (VariavelSubBase in subbase) {
    for (x in 2:9) {
      combinacao = combn(Motivos, x)
      MatrizGeral = c()
      NumCol = 0
      VariaveisAExplorar = c()
      for (VC in 1:length(combinacao[2, ])) {
        MatrizGeral = as.matrix(dados[noquote(combinacao[,VC])])
        NumCol = length(combinacao[,VC])
        VariaveisAExplorar = combinacao[,VC]
        VariaveisAExplorarExcel = c()
        VariaveisAExplorarSheet = c()
        for(u in 1:length(VariaveisAExplorar)) {
          VariaveiAExplorarNome =
          MotivosExcel[which(VariaveisAExplorar[u] == Motivos )]
          VariaveisAExplorarExcel =
          c(VariaveisAExplorarExcel, VariaveiAExplorarNome)
          VariaveisAExplorarSheetNome =
          MotivosExcelSheet[which(VariaveisAExplorar[u] == Motivos )]
          VariaveisAExplorarSheet =
          c(VariaveisAExplorarSheet, VariaveisAExplorarSheetNome)
        }

        Matrix = data.frame(MatrizGeral,
        dados[noquote(VariavelIndicador)])
```

```
DadosFim = data.frame(matrix(vector(), 3, NumCol,
           dimnames = list(c(), VariaveisAExplorar)),
           row.names = c("Importancia 1 e 2", "Importancia 3",
           "Importancia 4 e 5"))

for (variavel in VariaveisAExplorar) {
  numPrimeiro = 0
  numSegundo = 0
  numTerceiro = 0
  numFALSE = 0
  for (elemento in Matrix[, variavel]) {
    if (elemento == 1 || elemento == 2) {
      numPrimeiro = numPrimeiro + 1
    }
    if (elemento == 3) {
      numSegundo = numSegundo + 1
    }
    if (elemento == 4 || elemento == 5) {
      numTerceiro = numTerceiro + 1
    }
  }
  VectorAInserir = c(numPrimeiro, numSegundo, numTerceiro)
  DadosFim[variavel] = VectorAInserir
}

Ei = outer(rowSums(DadosFim), colSums(DadosFim),
"*")/sum(DadosFim)
NumAbaixo5 = length(Ei[Ei < 5])
NumAbaixo1 = length(Ei[Ei < 1])

if ((NumAbaixo5 < length(VariaveisAExplorar) * 4 * 0.2)
& (NumAbaixo1 == 0)) {
  P = "Eh possivel"
}

if (NumAbaixo5 > length(VariaveisAExplorar) * 4 * 0.2) {
  LinhasASair = unique(which(Ei < 5, arr.ind = T)[, 1])
  if (any(LinhasASair == 4)) {
    DadosFim = DadosFim[-4, ]
    Ei = outer(rowSums(DadosFim), colSums(DadosFim),
    "*")/sum(DadosFim)
    NumAbaixo1 = length(Ei[Ei < 1])
    NumAbaixo5 = length(Ei[Ei < 5])
    P = "Eh possivel"
    NumColunasNova = length(VariaveisAExplorar) * (nrow(Ei))
    * 0.2
    if (NumAbaixo1 > 0 || NumAbaixo5 > NumColunasNova) {
      P = "Nao eh possivel"
    }
  }
```

```
  if (4 %in% LinhasASair == FALSE) {
    P = "Nao eh possivel"
  }
}

if (P == "Eh possivel") {
  X2 = sum((DadosFim - Ei)^2/Ei)
  # estatistica do teste

  nu = prod(dim(Ei) - 1)
  # graus de liberdade

  pchisq(X2, df = nu, lower.tail = FALSE)
  # valor p do teste

  IndicadorSubBase = array(1:nrow(DadosFim))
  IndicadorSubBase[1] = VariavelSubBase
  IndicadorSubBase[2:nrow(DadosFim)] = ("")
  IndicadorSubBase = as.matrix(IndicadorSubBase)

  if (nrow(DadosFim) == 3) {
    ImportanciaAInserir = c("Importancia 1 e 2",
    "Importancia 3", "Importancia 4 e 5")
    ImportanciaAInserir = as.matrix(ImportanciaAInserir)
    SomaImportancia1E2 = sum(DadosFim[1,])
    SomaImportancia3 = sum(DadosFim[2,])
    SomaImportancia4E5 = sum(DadosFim[3,])
  # SomaImportanciaFALSE = 0
  }

  if (nrow(DadosFim) == 4) {
    ImportanciaAInserir = c("Importancia 1 e 2",
    "Importancia 3", "Importancia 4 e 5")
    ImportanciaAInserir = as.matrix(ImportanciaAInserir)
    SomaImportancia1E2 = sum(DadosFim[1,])
    SomaImportancia3 = sum(DadosFim[2,])
    SomaImportancia4E5 = sum(DadosFim[3,])
  # SomaImportanciaFALSE = sum(DadosFim[4,])
  }

  pvalue = pchisq(X2, df = nu, lower.tail = FALSE)

  if (pvalue >= 0.1) {
    PvalueDecisao = "Para os niveis usuais de significancia
    alfa, nao existe evidencia estatistica para rejeitar H0."
    RespostaAInserir = "Existe padrao de resposta!"
    Padrao = "SIM"
  }

  if ((pvalue >= 0.05) & (pvalue <= 0.1)) {
    PvalueDecisao = "Rejeito H0 para alfa maior que 5%."
```

```
  RespostaAInserir = "Ate ao nivel de 5% de significancia,
  existe padrao de resposta."
  Padrao = "SIM"
}

if ((pvalue >= 0.01) & (pvalue <= 0.05)) {
  PvalueDecisao = "Rejeito H0 para alfa maior que 1%."
  RespostaAInserir = "Ate ao nivel de 1% de significancia,
  existe padrao de resposta."
  Padrao = "NAO"
}

if (pvalue < 0.01) {
  PvalueDecisao = "Rejeito H0 para todos os
  niveis de significancia usuais."
  RespostaAInserir = "Nao existe padrao de resposta."
  Padrao = "NAO"
}

Pvalue = array(1:nrow(DadosFim))
Pvalue[1] = round(pvalue, 3)
Pvalue[2:nrow(DadosFim)] = ("")
Pvalue = as.matrix(Pvalue)

Decisao = array(1:nrow(DadosFim))
Decisao[1] = PvalueDecisao
Decisao[2:nrow(DadosFim)] = ("")
Decisao = as.matrix(Decisao)

Resposta = array(1:nrow(DadosFim))
Resposta[1] = RespostaAInserir
Resposta[2:nrow(DadosFim)] = ("")
Resposta = as.matrix(Resposta)

DadosFim = cbind(ImportanciaAInserir, DadosFim,
IndicadorSubBase, Pvalue, Decisao, Resposta)
MatrizExcel = data.frame(DadosFim)
NomesColunas = c("Importancia", VariaveisAExplorarExcel,
VariavelIndicador, "p-value", "Decisao", "Conclusao")
colnames(MatrizExcel) = NomesColunas

if (NumCol < 9) {
  JuntarVector = c()
  JuntarVector[1] = paste(JuntarVector,
  VariaveisAExplorarSheet[1])
  JuntarVector = stri_replace_all_fixed(JuntarVector, ' ', '')
  for (j in 2:length(VariaveisAExplorarSheet)) {
    JuntarVector = paste(JuntarVector,
    VariaveisAExplorarSheet[j])
  }
}
```

```
if (NumCol == 9) { JuntarVector = c('Todos') }

SubBaseExcelInserir = SubBaseExcel
[which(VariavelSubBase == subbase)]
NomeSheet1 = paste('{',SubBaseExcelInserir,'}')
NomeSheet1 = stri_replace_all_fixed(NomeSheet1, ' ', '')
NomeSheet = paste(JuntarVector, '+', NomeSheet1); NomeSheet

if (Padrao == "SIM") {
  ListaPvalueTabela = c(ListaPvalueTabela, pvalue)
  ListaPvalueTabela = as.matrix(ListaPvalueTabela)
  ListaMotivosTabela = c(ListaMotivosTabela, JuntarVector)
  ListaSubBaseTabela =
  c(ListaSubBaseTabela, SubBaseExcelInserir)
  ListaDecisao = c(ListaDecisao, PvalueDecisao)
  ListaImportancia12 =
  c(ListaImportancia12, SomaImportancia1E2)
  ListaImportancia3 =
  c(ListaImportancia3, SomaImportancia3)
  ListaImportancia45 =
  c(ListaImportancia45, SomaImportancia4E5)

  if (NumCol == 2) {
    write.xlsx(MatrizExcel, file = "Padroes - 2 Motivos.xlsx",
    sheetName = NomeSheet, col.names = TRUE,
    row.names = FALSE, append = TRUE)
  }
  if (NumCol == 3) {
    write.xlsx(MatrizExcel, file = "Padroes - 3 Motivos.xlsx",
    sheetName = NomeSheet, col.names = TRUE,
    row.names = FALSE, append = TRUE)
  }
  if (NumCol == 4) {
    write.xlsx(MatrizExcel, file = "Padroes - 4 Motivos.xlsx",
    sheetName = NomeSheet, col.names = TRUE,
    row.names = FALSE, append = TRUE)
  }
  if (NumCol == 5) {
    write.xlsx(MatrizExcel, file = "Padroes - 5 Motivos.xlsx",
    sheetName = NomeSheet, col.names = TRUE,
    row.names = FALSE, append = TRUE)
  }
  if (NumCol == 6) {
    write.xlsx(MatrizExcel, file = "Padroes - 6 Motivos.xlsx",
    sheetName = NomeSheet, col.names = TRUE,
    row.names = FALSE, append = TRUE)
  }
  if (NumCol == 7) {
    write.xlsx(MatrizExcel, file = "Padroes - 7 Motivos.xlsx",
    sheetName = NomeSheet, col.names = TRUE,
```

```
              row.names = FALSE, append = TRUE)
            }
            if (NumCol == 8) {
              write.xlsx(MatrizExcel, file = "Padroes - 8 Motivos.xlsx",
              sheetName = NomeSheet, col.names = TRUE,
              row.names = FALSE, append = TRUE)
            }
            if (NumCol == 9) {
              write.xlsx(MatrizExcel, file = "Padroes - 9 Motivos.xlsx",
              sheetName = NomeSheet, col.names = TRUE,
              row.names = FALSE, append = TRUE)
            }
          }
        }
      }
    }
  }
}

# CRIACAO DO PLOT

colors <- c('#4AC6B7', '#1972A4', '#965F8A', '#FF7070', '#C61951')

p = plot_ly(data , x = ~Importancia.4.e.5.., y = ~p.value,
color = ~Sub.Base, size = ~ni, colors = colors,
type = 'scatter', mode = 'markers', sizes = c(min(ni), max(ni)),
marker = list(symbol = 'circle', sizemode = 'diameter',
opacity = 0.7,
line = list(width = 2, color = '#FFFFFF')),
hoverinfo = "text",
text = ~paste('</br> Metrica:', Sub.Base,
              '</br> Motivos:', Motivos,
              '</br> Importancia 4 e 5:', Importancia.4.e.5..,'%',
              '</br> ni:', ni,
              '</br> p-value:', round(p.value,3),
              '</br> ni/total:',
              round(Quociente.Ni,2)*100,'%')) %>%

  layout(legend = list(font = list(size = 15)),
         title =  'P-value vs. Importancia',
         xaxis = list(title = 'Importancia 4 e 5 %',
                      range = c(min(Importancia.4.e.5..) * 0.9,
                      max(Importancia.4.e.5..) * 1.1),
                      gridwidth = 2),
         yaxis = list(title = 'p-value',
                      range = c(0,1.02),
                      gridwith = 2))

p
```

# B Criação de todos os modelos de Regressão Logística

```
dados = read_excel("ficheiro.xlsx")

Genero = factor(Genero);
EstadoCivil = factor(EstadoCivil);
EstadoCivil = relevel(EstadoCivil, "Casado")
Local = factor(Local);
TipoContrato = factor(TipoContrato)
DesempenhoGrupo = factor(DesempenhoGrupo)

Variaveis = c("Idade", "Antiguidade", "NumFilhos",
              "NumDependentes", "DistanciaKm", "Genero",
              "TipoContrato",
              "EstadoCivil", "GONum", "DesempenhoGrupo",
              "RetribuicaoTotalAnualPortugal",
              "MedianaRef", "MedianaGeral", "Target",
              "NumeroPromocoesSal", "AumentoSalarial",
              "HorizonteTemporalPromoSal",
              "NumeroMobilidades",
              "HorizonteTemporalMobilidades",
              "NumeroPromcoesFunc", "HorizonteTemporalPromoFunc",
              "TalentoFactor17")

NumeroVariaveis = ModeloInicial = FormulaModelo =
EficienciaMatrix = AccuracyMatrix = SensibilidadeMatrix =
EspecificidadeMatrix = AUCMatrix = CutOffMatrix =
VerdadeirosPositivosMatrix = VerdadeirosNegativosMatrix =
FalsosPositivosMatrix = FalsosNegativosMatrix = c()

for(x in 1:length(Variaveis)) { # 1 a Numero de Variaveis
  Combinacao = combn(Variaveis, x)
  for (y in 1:length(Combinacao[1,])) {
  # 1 a Numero de Combinacoes Possiveis
    VariaveisInserir = c()
    VariaveisNoModelo = c()
    CombinacaoEscolhida = Combinacao[,y]
    for (z in 1:x) {
      VariaveisNoModelo = ifelse(is.null(VariaveisNoModelo) ==
      "TRUE",
      paste(VariaveisNoModelo, noquote(CombinacaoEscolhida[z]),
      sep=""),
      paste(VariaveisNoModelo, "+",
      noquote(CombinacaoEscolhida[z]), sep=""))
    }
    VariaveisInserir = paste("Status4~",
    VariaveisNoModelo, sep = "")
    mod1 = glm(VariaveisInserir, family = binomial)

    # ********* DETETAR MULTICOLINEARIDADE **********
```

```
if(x > 1) {
  # Choose a VIF cutoff under which a variable is
  # retained (Zuur et al. 2010 MEE recommends 2)
  cutoffvif = 5
  # Create function to sequentially drop the variable
  with the largest VIF untilall variables have VIF > cutoff
  flag = TRUE
  viftable = data.frame()
  NumeroVariaveisVif = x

  while(flag==TRUE & NumeroVariaveisVif > 1) {
    vfit = vif(mod1)
    viftable = rbind.fill(viftable,as.data.frame(t(vfit)))
    if(max(vfit)>cutoffvif) {

      mod1 = update(mod1,as.formula(
      paste(".","~",".","-", names(which.max(vfit)))))
      NumeroVariaveisVif = length(all.vars(as.formula(mod1))) - 1
      }
    else {flag=FALSE}
  }
}

# ****** DEVIANCE *******
modelofull = mod1
modelonull = glm(Status4 ~ 1, family = binomial)

if(anova(modelonull,
modelofull, test = "Chisq")$"Pr(>Chi)"[2] > 0.05) next

# ******* ESCOLHA DO MELHOR MODELO *******
if(summary(modelofull)$df[1] - 1 > 1) {
  mod = step(modelofull,scope=list(lower=formula(modelonull),
  upper=formula(modelofull)), family =
  binomial,direction="both")

  if(formula(mod) != formula(modelofull)) {

    if(anova(modelofull, mod,
    test = "Chisq")$"Pr(>Chi)"[2] < 0.05) {
      modelofinal = mod
    }
    if(anova(modelofull, mod,
    test = "Chisq")$"Pr(>Chi)"[2] > 0.05) {
      modelofinal = modelofull
    }
  }

  if(formula(mod) == formula(modelofull)) {
    modelofinal = mod
  }
```

```
}

if(summary(modelofull)$df[1] - 1 == 1){
  modelofinal = modelofull
}

# ****** AVALIAR A SIGNIFICANCIA DOS COEFICIENTES - TESTE WALD *****

if(any(summary(modelofinal)$coeff[,4]) > 0.20) next

if(waldtest(modelofinal,
test = "Chisq")$"Pr(>Chisq)"[2] > 0.05) next

# ****** CAPACIDADE DESCRIMINATORIA DO MODELO *******

my_roc = ROC(form = formula(modelofinal),
             MI = FALSE, plot = 'ROC')

Soma = my_roc$`res`[,1] + my_roc$`res`[,2]; Maximo = max(Soma)
cutoff = my_roc$`res`[,5][which(Soma == Maximo)]
Sensibilidade = my_roc$`res`[,1][which(Soma == Maximo)]
Especificidade = my_roc$`res`[,2][which(Soma == Maximo)]

threshold = cutoff
predicted_values = ifelse(predict(modelofinal,
type="response")>threshold,1,0)
actual_values = Status4
conf_matrix = table(predicted_values, actual_values)

if(all(predicted_values) == 1) next

VerdadeirosPositivos = conf_matrix[2,2]
VerdadeirosNegativos = conf_matrix[1,1]
FalsosPositivos = conf_matrix[2,1]
FalsosNegativos = conf_matrix[1,2]

AUC = auc(Status4, predict(modelofinal,type="response"))

n = sum(conf_matrix)
accuracy = sum(diag(conf_matrix)) / n
eficiencia = (Sensibilidade + Especificidade)/2

NumeroVariaveisModelo =
length(all.vars(as.formula(modelofinal))) - 1

if (accuracy > 0.745 & eficiencia > 0.745) {

  NumeroVariaveis = as.matrix(c(NumeroVariaveis,
  NumeroVariaveisModelo))

  ModeloInicial = as.matrix(c(ModeloInicial,
```

```
      VariaveisInserir))

      if(x == 1) {FormulaModelo = as.matrix(c(FormulaModelo,
      VariaveisInserir))}
      if(x > 1) {
        VariaveisNoModeloFinal = c()
        for(w in 1:(NumeroVariaveisModelo+1)) {
          VariaveisNoModeloFinal =
          ifelse(is.null(VariaveisNoModeloFinal) == "TRUE",
                 paste(VariaveisNoModeloFinal,
                 noquote(all.vars(as.formula(modelofinal))[w]),
                 sep=""),paste(VariaveisNoModeloFinal, "+",
                 noquote(all.vars(as.formula(modelofinal))[w]),
                 sep=""))

        }
      FormulaModelo = as.matrix(c(FormulaModelo,
      VariaveisNoModeloFinal))
      }

      EficienciaMatrix = as.matrix(c(EficienciaMatrix, eficiencia))
      AccuracyMatrix = as.matrix(c(AccuracyMatrix, accuracy))
      SensibilidadeMatrix = as.matrix(c(SensibilidadeMatrix,
      Sensibilidade))
      EspecificidadeMatrix = as.matrix(c(EspecificidadeMatrix,
      Especificidade))
      AUCMatrix = as.matrix(c(AUCMatrix, AUC))
      CutOffMatrix = as.matrix(c(CutOffMatrix, cutoff))
      VerdadeirosPositivosMatrix =
      as.matrix(c(VerdadeirosPositivosMatrix, VerdadeirosPositivos))
      VerdadeirosNegativosMatrix =
      as.matrix(c(VerdadeirosNegativosMatrix, VerdadeirosNegativos))
      FalsosPositivosMatrix = as.matrix(c(FalsosPositivosMatrix,
      FalsosPositivos))
      FalsosNegativosMatrix = as.matrix(c(FalsosNegativosMatrix,
      FalsosNegativos))

    }
  }
}

Conclusoes = data.frame(NumeroVariaveis, ModeloInicial,
FormulaModelo, EficienciaMatrix,
                        AccuracyMatrix, SensibilidadeMatrix,
                        EspecificidadeMatrix, AUCMatrix,
                        CutOffMatrix,
                        VerdadeirosPositivosMatrix,
                        VerdadeirosNegativosMatrix,
                        FalsosPositivosMatrix,
                        FalsosNegativosMatrix)
Conclusoes
```

# Anexos

# A Template das Entrevistas de Saída

1. Nome do Colaborador:
2. Número de Colaborador:
3. Grupo Organizacional:
4. Avaliação de Desempenho (última):
5. Empresa:
6. Direcção:
7. Chefia Directa:
8. Local de Trabalho:
9. Descrição sucinta das funções desempenhadas e balanço global do percurso profissional na organização:
10. Qual o principal motivo que te levou a sair da Empresa?
11. Em que medida os seguintes factores contribuíram para a tua decisão de sair da Empresa? Escala: 1= contributo reduzido a 5 = enorme contributo.

Tabela 1: Fatores que contribuem para a decisão de sair da empresa.

| Factor | Importância | Comentários |
|---|---|---|
| Oportunidades de Desenvolvimento | | |
| Formação | | |
| Compensação | | |
| Benefícios | | |
| Avaliação de Desempenho | | |
| Relação com as Chefias | | |
| Equilíbrio entre a vida profissional e pessoal | | |
| Cultura Organizacional | | |
| Ambiente de Trabalho | | |

12. Outros factores que tenham contribuído para a saída? Quais?
13. Qual a empresa/sector para onde vais?
14. O que podemos fazer melhor para reter os nossos colaboradores?
15. Estarias receptivo(a) a voltar para a empresa no futuro?
16. Recomendarias a empresa como empregadora a um amigo?
17. O que mais valorizas nesta empresa?

18. O que deveríamos melhorar?
19. Entrevistado por:
20. Data: